Alma Mater Studiorum · Università di Bologna

Scuola di Scienze
Dipartimento di Fisica e Astronomia
Corso di Laurea in Fisica

Studies of CMS data access patterns with Machine Learning techniques

Relatore:
Prof. Daniele Bonacorsi

Presentata da:
Silvia De Luca

Sessione II
Anno Accademico 2015/2016

Contents

Sommario

Questa tesi presenta uno studio dei patterns di accesso allo storage su Grid in analisi distribuita da parte dell'esperimento CMS all'acceleratore LHC.

Questo studio spazia dall'analisi approfondita dei patterns di accesso ai file di CMS in passato (la cosiddetta "popularity"), fino all'utilizzo di un sistema di Supervised Machine Learning di tipo classificazione per prevedere i patterns di accesso ai dati dell'esperimento in futuro - con attenzione a particolari tipi di dati. L'esperimento CMS ha completato la sua prima fase di presa dati a LHC (Run-1) e, dopo un lungo periodo di shutdown (Long Shutdown 1) è iniziata nel 2015 la raccolta dei dati prodotti da collisioni protone-protone a 13 TeV del centro di massa nel periodo detto Run-2. I workflow di CMS vengono eseguiti su centri di calcolo (Tiers) di Worldwide LHC Computing Grid (WLCG), e in particolare l'analisi distribuita supporta le attività di centinaia di utenti al giorno. Le applicazioni CMS accedono a diversi tipi di dati ospitati sui sistemi di storage (disco) ai Tiers. Lo studio dettagliato di come questi dati vengono acceduti, in termini di tipologia, livello di Tier che li ospita e periodi di tempo in cui vengono acceduti permette di ottenere informazioni preziose sull'efficienza di utilizzo dello storage Grid di CMS, e in ultima analisi estrarre da esse suggerimenti per azioni concrete (ad esempio, pulizia delle cache disco e/o ulteriori repliche dei dati). In tal senso, l'applicazione di tecniche di Machine Learning permette di complementare quest'attività: l'apprendimento dei patterns di accesso a dati passati consente di costruire modelli con potenzialità predittive sugli accessi futuri.

Il **Capitolo 1** fornisce un'introduzione sulla Fisica delle Alte Energie e su LHC.

Il **Capitolo 2** descrive il modello di calcolo di CMS, con attenzione particolare al data management, introducendo anche il concetto di popolarità.

Il **Capitolo 3** descrive lo studio dei patterns di accesso ai dati di CMS con diverse "view" e livelli di approfondimento.

Il **Capitolo 4** offre un'introduzione a concetti di Machine Learning e spiega come vengono applicati in questo studio, descrivendo approcci seguiti e risultati ottenuti.

Abstract

This thesis presents a study of the Grid data access patterns in distributed analysis in the CMS experiment at the LHC accelerator.

This study ranges from the deep analysis of the historical patterns of access to the most relevant data types in CMS, to the exploitation of a supervised Machine Learning classification system to set-up a machinery able to eventually predict future data access patterns - i.e. the so-called dataset "popularity" of the CMS datasets on the Grid - with focus on specific data types.

The CMS experiment has completed its first data taking period at the LHC (Run-1) and, after a long shutdown (LS1), is now collecting proton-proton collisions data at 13 TeV of centre-of-mass energy in Run-2. All the CMS workflows run on the World-wide LHC Computing Grid (WCG) computing centers (Tiers), and in particular the distributed analysis systems sustains hundreds of users and applications submitted every day. These applications (or "jobs") access different data types hosted on disk storage systems at a large set of WLCG Tiers. The detailed study of how this data is accessed, in terms of data types, hosting Tiers, and different time periods, allows to gain precious insight on storage occupancy over time and different access patterns, and ultimately to extract suggested actions based on this information (e.g. targetted disk clean-up and/or data replication). In this sense, the application of Machine Learning techniques allows to learn from past data and to gain predictability potential for the future CMS data access patterns.

Chapter 1 provides an introduction to High Energy Physics at the LHC.

Chapter 2 describes the CMS Computing Model, with special focus on the data management sector, also discussing the concept of dataset popularity.

Chapter 3 describes the study of CMS data access patterns with different depth levels.

Chapter 4 offers a brief introduction to basic machine learning concepts and gives an introduction to its application in CMS and discuss the results obtained by using this approach in the context of this thesis.

Chapter 1

High Energy Physics at the LHC

1.1 The Standard Model

The Standard Model [3][4], developed in 1970, incapsulates the remarkable theories and discoveries of particles physics since 1930, it has an important feature: it is verified by all available data, secondly it gives a unified description in terms of quantum field of all the interaction of known particles (except gravity): Electroweak and Quantum Chromodynamic; Gravitation si excluded because SM hardly faces phenomena at Planck's scale of energy $10^{-19}GeV$ (of interest in cosmology). The SM also describes all particles and splits them into two main classes according to their intrinsic angular momentum: *fermions*, that have half-integer spin, and *bosons* which have integer spin. There are twelve fermions respectively six *leptons* and six *quarks*, the building blocks of matter. Moreover, for every particle there is its own antiparticle: a particle that differs only for opposite internal quantum numbers. Quarks can have three different colors (red, blue, green)and only mix in such ways as to form colourness objects, while lepton have either unitary or null electric charge and they can be organized in three generations (with associated neutrino):

$$\begin{pmatrix} \nu_e & \nu_\mu & \nu_\tau \\ e & \mu & \tau \end{pmatrix}$$

Electron, muon, tau (second row) and their associated neutrino (first row) but there is a phenomena by which neutrinos can evolve into a different kind i.e. ν_e becomes ν_μ and so on. Electron, muon and tau interact by both electromagnetic and weak force whereas neutrinos only by weak force. Quarks interact by strong force that is regulated by colour charge. We have in this model a very detailed description of fundamental interaction: the strong force, the weak force, the elctromagnetic force and the gravitational force. They work over different ranges and have different strenghts: gravity is the weakest but it has and infinite range; the electromagnetic force also has infinite range but is many time

stronger than gravity. The weak and strong forces are effective only over a very short range and dominate only at the level of subatomic particles. Three of the fundamental forces result from the exchange of force-carrier particles, called bosons. Each fundamenta force has its own boson: the strong force is carried by the *gluon*; the electromagnetic force is carried by the *photon*; the interaction of weak force is due to the W and Z bosons; the corrisponding force-carrying particle of gravity should be the *graviton* but not yet found.

There are also questions that it does not answer regarding to: what is dark matter, what happened to the antimatter after big bang, how it is possible that there are three generations of quarks and leptons with such differen masses, etc. A great goal achived was the confirmation of Higgs boson existence, an essential component of the Standard Model.

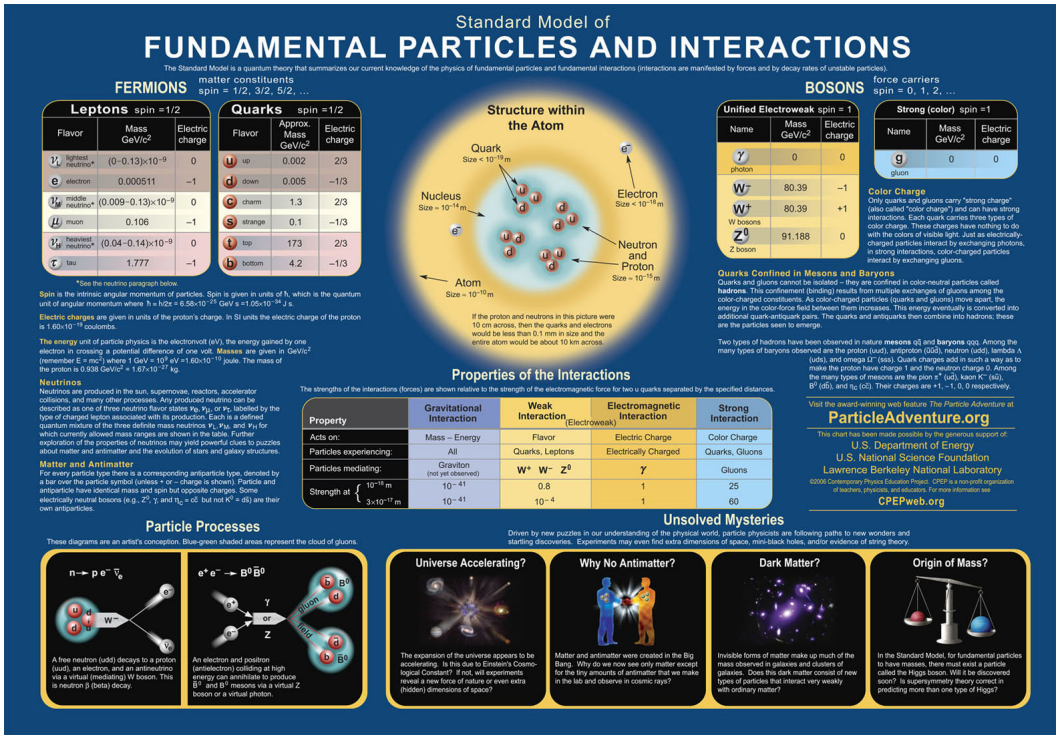

Figure 1.1: summary of Standard Model basic concepts.

1.2 CERN and the LHC

The European Organization for Nuclear Research, known as CERN [1][2] is a research organization that operates the largest particle physics laboratory in the world; the term CERN is also used to refer to the laboratory. In 1954, 12 European nations came together to sign the convention officially forming CERN and nowadays has 22 member states. It is concived to study the basic costituents of matter - the fundamental particles; a number of clues about how the particles interact and insights into the fundamental laws of nature came from collisions between particles (close to the speed of light). At CERN purpose-built particle accelerators and detectors are used to investigate. Accelerators boost beams of particles to high energies before the beams are made to collide with each other or with stationary targets. Detectors observe and record the results of these collisions. Initially CERN was focussed of pure physics research ad understanding the inside of the atom. Today we go beyond "nuclear" world and aim to explore the particle physics (the fundamental costiuents of matter and the forces between them) and much more ranging from high-energy phisics, from studies of antimatter to the possible effects of cosmic rays on clouds. The particle physics have described the fundamental structure of matter using the Standard Model: describes how everything that we observe in the universe is made from a few basic blocks called fundamental particles, governed by four forces. Then, at CERN accelerators are used also to test the predictions of standard model. One takes account that the model only describes the 4% of the known universe. CERN is not only particle physics, one remarcable thing happened in 1989 when Tim Berners-Lee, a British scientist, invented the World Wide Web (WWW); its initial purpose was to meet the demand for automatic information-sharing between scientists in universitied and institutes around the world. The first web site described the basic features of the web; the software of World Wide Web was put in the public domain in 1993. Other physics topics are for example: compositeness, the high energy collisions at LHC could be the key to find a possible substructure for subatomic particles; cosmic rays that are rays of charged particles which energy is far higher than LHC's; dark matter, a misterious matter that makes up most of the universe; extra dimensions, gravitons; heavy ions and quark-gluon plasma to recreate similiar condition of universe just after the Big Bang.

The Large Hadron Collider [5][6] is the world's largest and most powerful particle accelerator. It first started up on September 2008, and stands as the main component of the accelerator complex. The LHC, built in a tunnel buried 175m unerground, consists of a 27 kilometer ring of superconducting magnets with a number of accellerating structure to boost the energy of the particles along the way. Inside the accelerator, two high-energy particle beams travel at close to the speed of light before they collide. The beams are stored for hours; during this time collisions take place inside the four main LHC experiments. The beams travel in opposite directions in separate beam pipes - two tubes kept at ultrahigh vacuum. They are guided around the accelerator ring by a strong magnetic

field maintained by superconducting electromagnets. The electromagnets are built of coils that operate in a superconducting state, efficiently conducting electricity without resistance or loss of energy. This requires a connection to huge cryogenic stystems which cool the magnets at $-271.3C$.

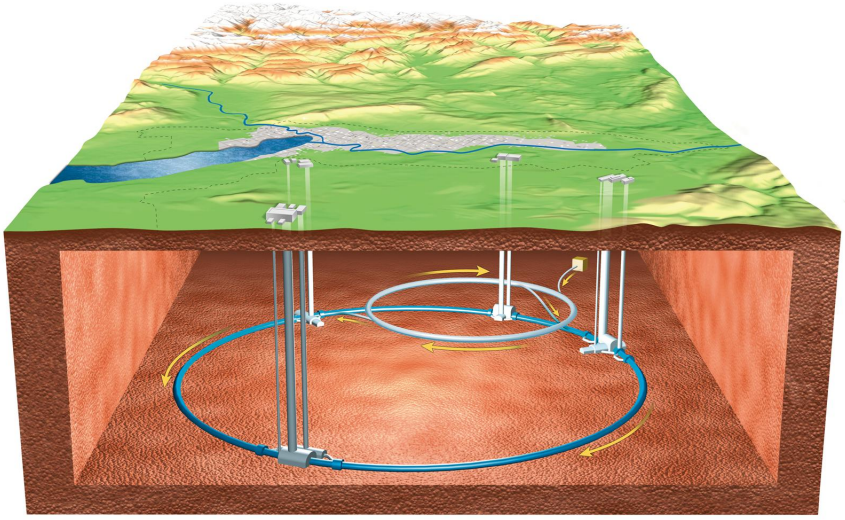

Figure 1.2: Graphic representation of LHC tunnel.

The LHC is a proton-proton and heavy ion collider. At the start of the accelerator's complex there is a bottle full of gaseous hydrogen from which the ionization process starts using a duoplasmatron to generate protons. The idea of a circular accelerator arises from the need to work at high energy level. Particle acceleration process is made possible by and electric field generated by a system of electrods alternating poles. At each step, the proton increases its velocity therefore the time interval from previous step to the next is each time shorter. A solution is the linear accelerator model LINAC [7]. In real application, the acceleration process is performed by using resonant cavities and radio-frequency generators. Thus with a LINAC could be hard reaching energies about TeV, because it entails a longer accelerator. To avoid this incovenient, scientists thought about a "ring" structure composed by linear accelerator steps (LINAC) connected to each other by magnetic dipole to bend proton's trajectory and to keep them in line. The LHC dimensions are aimed at minimizin the loose of energy under synchrotron radiation which depends on the bending ray.

Protons are allowed to enter the LHC accelerator [6] when they reach a 450 GeV energy. To reach this target there are four pre-accelerating steps represented in Fig.1.3:

- **LINAC2** (1978,36m in circumference): from the bottles with gasseous hydrogen to 50MeV energy;

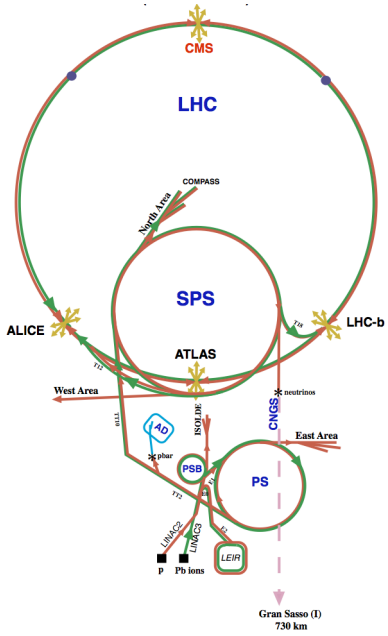

Figure 1.3: Schematic representation of accelerators complex

- **PSB** (Proton Synchrotron Booster,1972,157m in circumference): from LINAC2 to 1,4GeV energy;
- **PS** (Proton Synchrotron): from 1,4GeV to 25GeV;
- **SPS** (Super Proton Synchrotron,1978,36m in circumference): from 25GeV to 45GeV;
- **LHC** (Large Hadron Collider): to theoretical 14TeV ;

Protons quit SPS from two different injection points. Most of LHC ring is composed by magnets, there are text two different circles (with distinct magnetic fields) with the purpose to keep protons running in clockwise and counterclockwise. Those two parallel rings cross into only four points namely where the main experiments are located. In more detail, as already mentioned, the layout of straight sections depends on the specific use of the insertion, for example:

- physics events such as beam collision;
- injection;
- beam dumping etc.

One of LHC characteristic machinery is its Vacuum System [8], as collisions against gas molecules must be avoided. LHC gets three of those: insulation vacuum for cryomagnets, insulation vacuum for helium distribution line (QRL), beam vacuum. At cryogenic

temperatures, in the absence of any significan leak, the pressure will be stabilised aroud $10^{-4}Pa$. The requirements for the beam vacuum are much more stringent, driven by the requested beam lifetime and background to the experiments. The requirements at cryogenic temperature are expressed as gas densities and normalised to hydrogen, should remain below $10^{15}H_2m^{-3}$ to ensure the required 100 hours beam lifetime. All three vacuum system are subdivided into sectors by vacuum barriers for the insulation vacuum and sector valves for the beam vacuum. The beam vacuum is divided in sectors of various lengths. A number of dynamic phenomena have to be taken into account for the design of the vacuum system for example synchrotron radiation and electron clouds. A crucial task of experiments in LHC is to make collisions detections. The beams are made up of proton bunches and each one can circulate for many times during the same run; the experiments' detectors are synchronized along the collisions through a clock whose fundamental frequency coincides with bunches' position within their trajectory. The protons distribution is choosen to make collision exactly in the middle of detector, however not all protons collide at once so to optimize significant events, one can do basically three things: squeezing bunches is needed to make bunches more compact and aligne the beams; encrease the number of protons in a bunch; raise the number of bunches per run. Each proton beam at full intensity will consist of 2808 bunches; each bunch will contain

Particles accelerated	Protons and heavy ions (Lead 82+)
Accelerator circumference	$26659m$
Injected beam energy	$450GeV$ (protons)
Nominal beam energy for physics	$7TeV$ (protons)
Magnetic field at $7TeV$	$8.4T$
Operating temperature	$1.9K$
Number of magnets	1232
Number of quadrupoles	858
Number of correcting magnets	6208
Number of RF cavities	16
Frequency of RF cavities	$400MHz$
Maximum Voltage of a single RF	$2MV$
Maximum Luminosity	$\mathcal{L} = 10^{34}cm^{-2}s^{-1}$
Power consumption	$\sim 180MW$

Figure 1.4: Main parameters of the LHC.

$1.15 * 10^{11}$ protons at the start of nominal fill. Total beam energy at the maximum is 352MJ. The bunches are generally far about 25ns from each other; however there are some holes in the bunch structure, the biggest is the beam abort gap of $3\mu s$. This is there to give to the beam bump kickers time to get up to full voltage. There are also other smaller gaps in the beam which arise from similiar needs from the SPS and LHC injection kickers.

1.3 Experiments at the LHC

Four major experiments at the LHC use detectors to analyse a multitude of particle producted during collisions. The largest particle detectors are those used by the ATLAS and CMS collaborations, to explore a large variety of phenomena at the highest LHC energy scales. ATLAS and CMS, along with other experiments operating at the LHC, are briefly presented in the following.

ALICE (A Large Ion Collider Experiment) [9][10] is a detector designed for heavy-ion collision. It is built to study the physics of strongly interacting matter at extreme energy densities, where a phase of matter called quark-gluon plasma is formed. The quarks, as well as the gluons, seem to be bounded permanently together and confined inside composite particles, such as protons and neutrons. Collisions in the LHC are such hot that recreate in laboratory condition similiar to those just after the big bang. Under these conditions, protons and neutrons melt, freeing the quarks from their bonds with the gluons, thus creating the quark-gluon plasma. The existence and properties of that phase are key issues in the theory of quantum chromodynamics (QCD), understanding the phenomenon of confinement, etc. ALICE studies such state of matter as it expands and cools, and how it ultimately gives rise to the particles we actually observe. The ALICE detector is 26m long, 16m high and 16m wide.

ATLAS [11][12] is the other LHC general-purpose detector; beams of particles from the LHC collide at its detector centre producing debris as new particles, which flow out of the collision point in all directions. Six different detecting subsystems arranged in layers around the collision point record the trajectory, momentum, and energy of the particles. A huge magnet system bends the tracks of charged particles so their momenta can be measured. ATLAS uses a trigger system to tell the detector which events to record and which to reject. Data-acquisition and computing systems are used to collect, handle and analyse the collision events recorded. The ATLAS detector is 46m long, 25m high and 25m wide.

CMS (Compact Muon Solenoid) [13] has a broad physics scope ranging from in-depth studies of the Standard Model to searching for new physics. ATLAS and CMS share the same physics goals, but attack them with detector of quite different conception. The CMS detector layout is designed around a huge solenoid, a cylindrical coil of superconducting cable that generates a field of 4 tesla and this is confined by a steel yoke. The CMS detector is 21m long, 15m wide and 15m high.

LHCb (Large Hadron Collider Beauty) [14][15] is specialised in investigating the slight differences between matter and antimatter by studying the quark b (beauty). LHCb uses a series of subdetectors to mainly detect forward particles. LHCb is

composed by a sophisticated spectrometer and planar detectors; it is 21m long, 10m high and 13m wide.

LHCf (Large Hadron Collider Forward) [16] uses forwards particles thrown by collisions in the LHC as a source to simulate cosmic rays in laboratory conditions. Cosmic rays are a natural source of charged particles: while colliding with high atmosphere, a cascade of particle is produced and reaches on Earth's surface. LHCf consists of two detectors which sit along the LHC beamline, at either side of the ATLAS collision point: this location allows the observation of particles at nearly zero degrees to proton beam direction. Each detector is 30cm long 80cm high and 10cm wide.

TOTEM (Total elastic and diffractive cross-section measurement) [18][19] is about taking measurements of protons as they emerge from collisions at small angles, a region known as forward direction, and is inaccessable by other LHC experiments. It is equiped of four particle telescopes and 26 Roman pot detectors, spread around the CMS interaction point. The telescopes use cathode-strip chambers and Gas Electron Multipliers to track the particles emerging from CMS collisions. Roman Pots with silicon sensors perform measurements of scattered protons.

MOEDAL (Monopole and Exotics Detector at the LHC) [20] searches magnetic monopole, a hypothetical particles with a magnetic charge. This monopole detector is an array of 400 modules, each consisting of a stack of 10 sheets of plastic nuclear-track detectors. This detector is deployed around the same intersection regions as the LHCb detector. The magnetic monopoles, if they exist, would rip through the detector, breaking long-chain molecules in the plastic nuclear-track and creating a minute trail of damage through all 10 sheets. Another research is for highly ionizing stable massive particles, predicted by the Standard Model.

Chapter 2

The CMS Experiment

2.1 The CMS detector

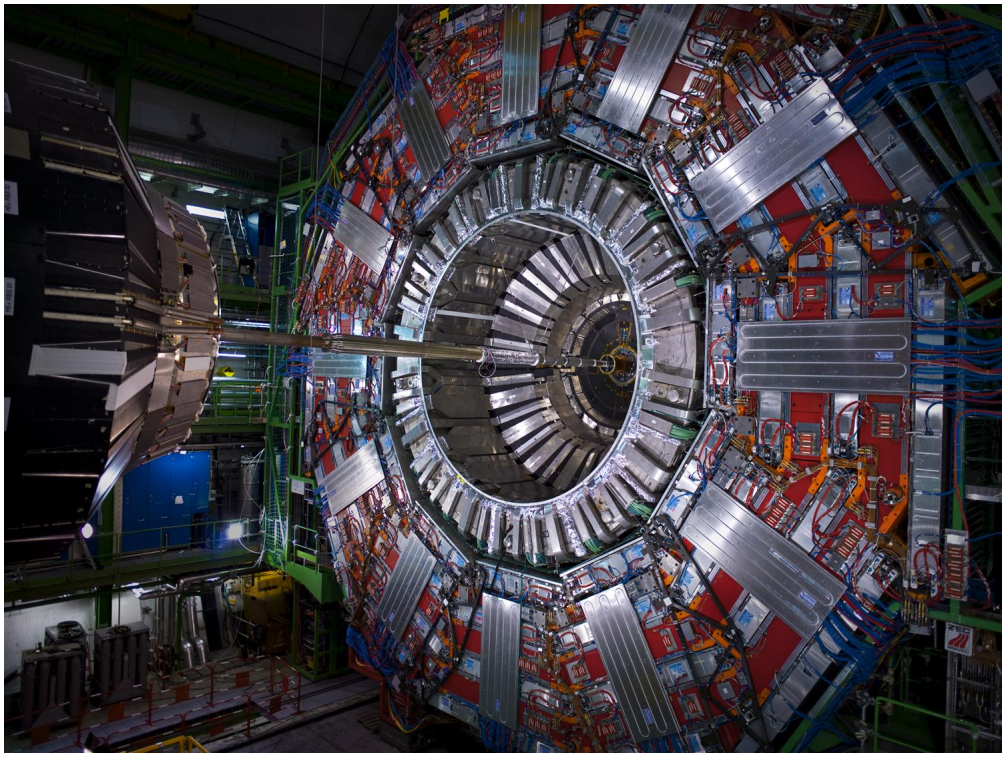

Figure 2.1: Picture of the CMS detector while open.

The Compact Muon Solenoid detector [21] works at the LHC at CERN. It is a multi-purpose detector designed to proton-proton (and lead-lead) collisions. Main CMS work-parameters are: *centre of mass energy* (14Tev) and *luminosity* $10^{34}cm^{-2}s^{-1}$. The

CMS detector is a stratified fashioned structure consists of five layers: the tracker, the electromagnetic calorimeter, the magnet, the hadronic calorimeter, and the muon system. This is planned to stop, track or measure a different kind of particle derived from early collision. The detector uses a powerful solenoid that bends the trajectory of charged particles. Data given by the detector are stored and then used to recreate what happens at the core of the collision; to do so a *Trigger* and a *Data Acquisition System* are needed. When collisions at maximum designed energy occur, 10^9 events per second will result. So, a significant number of events take place but the on-line selection process has to trigger only $100/s$ will be saved. This makes and outstanding necessity for a custom electronics capable both to manage high data flux and being able to make extremely careful selections.

The detector design addresses to those general requirements:

- A high performance system to identify and track muons;
- A high resolution electromagnetic calorimeter to detect and measure elctrons, positrons and photons;
- A high quality tracking system for momenta measurements;
- A hermetic hadron calorimeter, designed to entirely surround the collision and prevent particles escaping;

Among the stored data there are the momentum and energy of particles; then, by their combination one know what type of particle is and, by tracing back patterns, at least its mass.

It follows a brief description of each system.

Tracker

The Tracker [22] is the innermost element of the detector and its task is to detect muons, electrons, hadrons and particles coming from decay. This is made entirely of silicon-based technologies. Particles leave traces of energy that allow to chart their flight paths (position) which are spiral shaped and their curvature reveals their momenta. During momentum measurement, the interaction between the tracker and the particles must be least as possible. The tracker is endowed with two types of technologies for particle detection: pixels and miscro-strips. Outside the detector the signals are transferred through optic fibres cables. Moreover, a second measurement of momentum is performed by an outer muon chamber system, it allows to have a reconstruction mean efficiency about 92%.

Electromagnetic Calorimeter: ECAL

The CMS electromagnetic calorimeter [23] plays a leading role studying the electroweak symmetry breaking and detection of two-proton decay and of electrons

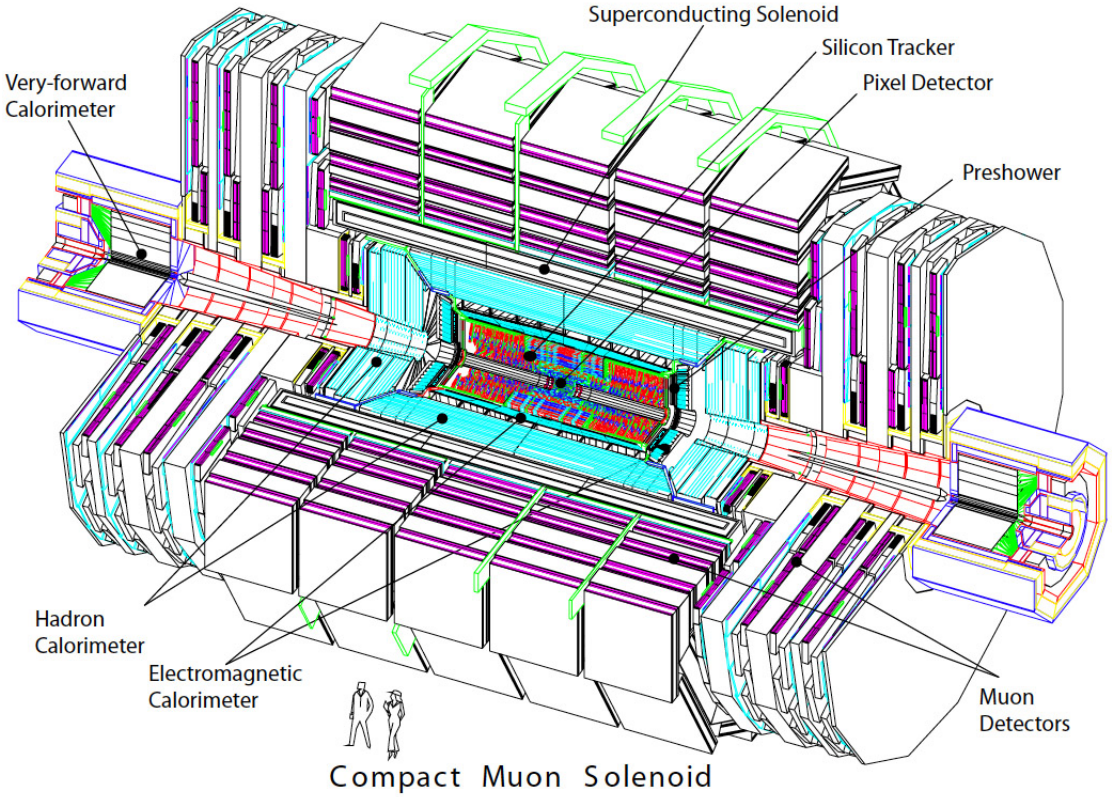

Figure 2.2: Section of the CMS detector.

and positrons coming from Ws and Zs decay. It is main based on scintillating tungstate $PbWO_4$ crystals structure that covers the entire solid angle, and offers advantageous features:

- It is a fast scintillator (fast light emission);
- High performance energy resolution;
- Short radiation leght;
- Easy production process from raw materials;

The light produced by particles hitting the tungstate crystals must be recorded. This operation is performed by Avalanche Photo Diodes (APDs) places around the calorimeter (barrel region) and by Vacuum Photo Triodes (VPTs) at the endcaps. In front of the endcaps there is a preshower detector made of two lead silicon detector layers to distinguish single high-energy photons from pairs of low-energy photons. The photons are detected by a sensor and it is possible to estimate their initial energy.

Hadron Calorimeter: HCAL

The combined CMS calorimeter system examines the direction and energy of standard model particles. The Hadron Calorimeter [24] measures mainly hadron jets, neutrinos (as missing energy transverse) and also cooperates with ECAL and muon detector in the identification of electrons, protons and muons. HCAL must be a hermetic structure, that is make sure it captures every particle emerging from collisions. HACL is a sampling calorimeter meaning it finds a particle's position, energy and arrival time using alternating layers of absorber and fluorescent scintillator materials that produces a light pulse if bumped by a particle. A system of optical fibres collects up this light and sends it into a readout boxes were photodetectors amplify the signal. The total amount of light obtained by summing up the light measured is a valuation of particle's energy. The HCAL is organised into barrel (HB and HO), endcap (HE) and forward section (HF).

Magnet

The CMS magnet is the central device around which the experiment is built, with a 4 Tesla magnetic field. This is a solenoid of superconducting material, a magnet made of coils of wire that produce a uniform magnetic field when electricity (CMS uses 19500 A) flows through them. The Tracker and the calorimeters (ECAL, HCAL) fit inside the magnet coil whilst the muon detectors are interleaved with a 12-sided iron structure that surrounds the magnet coil and contains and guides the field. This is a return yoke made up of three layers also acts as a filter, allowing through only muons and weakly interacting particles.

Muon Detector

Detecting muons [25] is one of CMS's most important tasks, they are produced in the decay of a number of potential new particles. Muons are relatively non-interacting particles; they can penetrate several meters of iron without interacting, then they cannot be stopped by any CMS's calorimeter. Therefore, chambers to detect muons are placed at the very edge of the experiment. There are three types of subdetectors for muons' identification. The particle's measuring process begins by fitting a curve to hits among the four muon stations (detectors), which sit outside the magnet coil and are interleaved with return yoke plates. By tacking its position through the multiple layers of each station, combined with tracker measurements the detectors precisely trace a particle's path. In total there are 1400 muon chambers, 250 drift tubes (DTs) and 540 cathode strip chambers (CSCs) to trace particle's positions and give a trigger, 610 resistive plate chambers (RPCs) form a trigger system to select data acquired. DTs and RPCs are arranged in concentric cylinder around the beam line, while CSCs make up the endcaps disks at both ends of the barrel.

Trigger

The Trigger [26] system was added to manage the data when CMS is performing at its peak: about one billion inelastic proton-proton collisions take place every second. A slice of those data couldn't be useful, then the Trigger and Data Acquisition System operated a first selection so those data can be stored with a reduced rate. The Event selection is divided in two stages:

- *Level-1 Trigger*: at this level the events stored rate is no more than 100kHz and are forwarded to High Level Triggers. The L1 Trigger is organised into three subsystems and based on costum electronics (ASICs and FPGAs): the L1 calorimeter trigger, the L1 muon trigger and the L1 global trigger. To perform the event selection the trigger system has a determined time lapse, $3\mu s$ after each collision then data temorally saved in the buffer are overwritten.
- *High Level Trigger*: it relies on software implementation and it is the next step in the event selection made by L1 Trigger. Each processor is connected, by design to all the detector elements of CMS, and can therefore access any data it deems valuable for the selection of any particular event, whose set is the HLT harwdare. The HLT firstly evaluates a L1 candidate and continues the L1 reconstruction; then, if the candidate is stored, it reconstructs its tracks using also the tracker's information. This operation is very CPU-expansive, thus not all parts are reconstructed, only strictly required ones [27][28].

2.2 Software and Computing in CMS

2.2.1 Worldwide LHC Computing Grid

Figure 2.3: View of the CERN Computing center.

The Worldwide LHC Computing Grid project [29][30] coordinates the deployment and operations of computing centres used for LHC activities. On the WLCG resources, the LHC experiments store the data and perform processing tasks. The amount of data collected per year by all LHC experiment reaches the scale of tens of Petabytes and the amount of processing power need sums up to tens of millions of jobs. So, it was not conceivable to design and build a huge computing center in one nation, but building a worldwide infrastructure of computing centers strongly interconnected was the only option. WLCG today connects the computing resources of 170 centres spread in 41 countries all operating Grid middleware. The WLCG relies mainly on two Grids: the European Grid infrastructure [31] and the Open Science Grid (USA) [32]. The computing centers included in WLCG are hierarchically organised in "Tiers". This project provides crucial features to face LHC challenges [33]:

- A faster access to resources by making multiple copies of data kept at different sites;
- Data equally available indipendent of users' location;
- Computer centers in multiple time zones ease round-the-clock monitoring and expert support;
- Resources can be distributed across the world, for funding and sociological reasons.

In the HEP community (and not only), the Grid services and middleware are intended to be usable by more experiments as Virtual Organizations (VO) [34]. On top of the

common middleware layer, each VO can pass its own experiment-specific application layer. The elements which constitute every Grid site are:

Computing Element manages the jobs submitted by the user and the interactions with the Grid services.

Worker Node is where the computation actually happens.

Storage Element allows access to storage and data at site. Data can be stored on different type of storage resources: *tapes* are used as long-term storage media, whereas in *disks* are used for more performant data access.

User Interface is the resource on which a user enters into the Grid.

Central Services are a set of services running centrally which are needed for workload and data management such as: data catalogues, workload management systems, and data transfer solutions.

Another important storage infrastructure is the Storage Federation. It provides access to the data on Storage Element, but it doesn't rely on a catalogue. It uses a set of so-called "re-directors" that - once a file as been search on local storage and eventually not found - redirect the data access request to another location, browsing higher in the data organization and finding a data access location.

As stated previously the computing centres are organised in four Tier levels (see Figure 2.4):

- **Tier-0** is deployed in two locations: CERN Data Centre in Geneva and Wigner Research Centre for Physics in Budapest. Tier-0 is responsible for the safe-keeping of the RAW data coming out of the detector, for first pass reconstruction, distribution of raw data and reconstruction output to the Tier-1s, and contributed to other tasks like data reprocessing.

- **Tier-1** these are 13 computer centres with sufficient storage capacity but only 2 of them serve the needs of just one LHC experiment, all the others support more than one LHC experiment. Tier-1s operate a safe-keeping of a proportional share of RAW and RECO data, large-scale reporcessing and safe-keeping of corresponding output, data distribution to Tier-2s and sake-keeping of a share of simulated data produces at these Tier-2s.

- **Tier-2** there are about 160 Tier-2s in LHC, usually not CERN-based such as universities or scientific institutes. They provide CPU and storage for processing, both Monte Carlo productions and data analysis, with a balanced share among the two - but they are the primary resource for data analysis in the most experiments. They

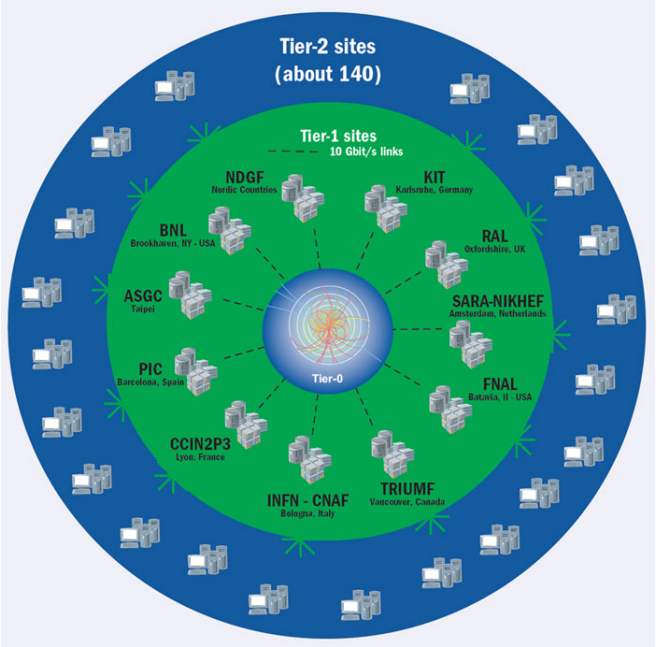

Figure 2.4: Graphical representation of WLCG Tier levels.

handle analysis requirements and proportional share of simulated data productiona and reconstruction but do not have tape archiving.

- **Tier-3** is a local computing resources for the users which can consists of local computer clusters or even just an individual PC. There is no formal engagement between WLCG and Tier-3 resources.

2.2.2 CMS Computing Model

The CMS project faces challenges not only in terms of the detector operation and the physics program, but also in terms of the data handing on computing resources. Most CMS collaborators are not CERN-based, and have access to significant non-CERN resources; therefore, the CMS computing environment [36] is based upon Grid middleware, with the common Grid services at centres defined and managed through the WLCG as explained beofore. Then, this computational infrastructure is intended to be available to CMS collaborators wherever they are. Basically a system of this kind must fulfill those needs:

- The analysis of a very large dataset requires a system of large scale, with an efficent approach to data reduction and patter recognition.

- Highly flexibility to make easy access to any data within the lifetime of the experiment. It must supports a wide variety of data processing tasks.
- A primary characteristic should be the manageability in its functionalities (i.e. computing operation and software thing)

Key components of the computing system [37] include:

- An event data model and correspondig application framework;
- Distributed database system;
- A set of computing services;
- Underlying generic Grid services giving access to distributed computing resources;
- Computer centres, managing and providing access to storage and CPU at local level.

The disign challenges have been addressed through construction of a modular system of loosely coupled components with well-defined interfaces, and with emphasis on scalability to very large event samples.

The CMS application sofware must perform a variety of event processing, selection and analysis tasks, both online and offline. The central unit of the CMS data is the *Event*, which corresponds to a single bunch crossing; subsequent events with constant instantaneous luminosity form the *Lumisections*, and several Lumisections form the *Run*. Events from Runs are arranged into *Datasets*. Some data formats, with respect to their properties, are based on the Event: RAW digitised data, reconstructed produtc etc. The Event also contains information describing the origin of the RAW data and the provenance of derived data. The latter information allows users to unambiguously identify how each event contributing to a final analysis was produced; it includes a record of the software configuration and conditions/calibration setup used to produce each new data product. Events are physically stored as persistent ROOT files.

CMS makes use of several formats with different levels of detail of the information contained. The process of data reduction and transformation among formats takes place in several steps, usually is carried out at different computer centres.

RAW

RAW events contain the full recorded informations from the detector, plus a record of other metadata. RAW data is accepted into the offline system at the HLT output rate, and extension of the RAW data format is used to store the output of CMS Monte Carto simulation tools. This data type is permanently archived in safe storage and occupies roughly 1.5MB/event. The RAW data are classified by the online system into

CMS data tier	Size per event (MB)
RAW Data	0.9
MiniAOD Data	0.032
RECO Data	2.2
AOD Data	0.370
RAW MC	1.5
MiniAOD MC	0.380
RECO MC	2.4
AOD MC	0.41

Table 2.1: Size of selected CMS data tiers, for Data and MC, at expected average PU 35.

several distinct *primary datasets*. This classification has several advantages such as the possibility of assigning priorities to data reconstruction and transfer. CMS will also define some flexible "express streams" used for prompt calibration and rapid access to interesting or peculiar events.

RECO

Reconstructed (RECO) data are produced by applying several levels of patter recognition and compression algorithms to the RAW data. RECO data contains objects created by event reconstruction that is the most CPU-intensive activity in the CMS data processing chain and is made by mainly four steps:

1. Detector-specific processing (data decoding, application of detector calibration constants and objects are reconstructed).

2. Tracking (include reconstruction of global track from hits in the silicon and muon detectors).

3. Vertexing (reconstruction of primary and secondary vertex candidates)

4. Particle identification (produces the standard physics object which is most matching with physics analyses).

The RECO are not permanent in fact they can be recalculated when newer calibrations and better software are available. RECO events contain both the low level physics objects as hits, clusters etc and the high level objects such jets, muons, electrons etc. Moreover there is a direct connection between high and low level objects that avoids duplication of information. RECO events occupy roughly 0.5MB/event.

AOD

Analysis Object Data (AOD) is the compact analysis format, designed to be relatively small in size thus allowing large data samples in such format to be hosted in many computing centres. AOD events contain high-level physics objects, and additional information to allow kinematic refitting. AOD data are produced by filtering RECO data, either in bulk production or in a skimming process which may also filter a primary dataset into several analysis datasets.

AODSIM

AODSIM events are produced through Monte Carlo methods. They contain high-level information and are used for physics analyses.

MiniAOD

MiniAOD is a data tier of CMS data introduced in Spring 2014 to serve the needs of the mainstream physics analyses while keeping a small event size (30-50 kb/Event). The main contents of the MiniAOD are: high-level physics objects with detailed information; the full list of particles reconstructed by the ParticleFlow; trigger information, MiniAOD contains the trigger bits associated to all paths; plus there are objects reconstructed at L1 and the L1 global trigger summary [38].

In addition to event data, there is a variety of *Non-Event Data*, that is required in order to interpret and reconstruct events. CMS uses 4 types of Non-Event data: construction data, equipment management data, configuration and conditions data. Non-Event data are held in central Oracle databases, for access by online and offline applications. Conditions data access at remote sites take place via FroNTier system [39] which uses a distributed network of cache http proxy servers.

Within the CMS Computing Model 2 main sectors can be identified: the *CMS Data Management System*, focussing on access and storage aspects, and the *CMS Workload Management System* focussing on jobs flux handling. The detailed description of all the details of both sectors goes beyond the scope of this thesis. On the other hand, some aspects are important to set the context for the next chapters, so a brief and not exhaustive description is provided in the following.

2.2.3 Data and Workflow management

The CMS data management sector covers the cataloguing or alternative solutions to track the location of physical files on site storage systems, the transfer of files across sites, the aspects of data access, etc. The information about which data exists is offered by the CMS *Dataset Bookkeping System*. The Data Bookkeping Service [41] provides access to a catalogue of all event data, both from Monte Carlo data and collisions data, and it records the files metadata including its processing derivation (i.e. with the information in the catalogue it is possible to track back to the original RAW data or Monte Carlo generation). The data managed by DBS are not actually data but pointers to physical data organised by other parts of CMS data management system. The dataset replication system, i.e. static data placement, is implemented by the PhEDEx system. The PhEDEx (Physics Experiment Data Export) [42] is a reliable and scalable dataset replication system targeted to serve large scale data transfer needs across Grids. PhEDEx provides a centralized system for making global data movement decisions and a realtime view of the global CMS data transfer state. PhEDEx is composed of a seried of autonomous, persistent processes, called agents, which share information about replica and transfer state through a database. Agents are very specialized, e.g. there are agents for download, for tape migrations, etc. Recently, Dynamic Data Management features have been added to dynamically delete unused data replicas and further replicate heavily accesses ones.

The workload management sector of the CMS Computing Model is focused on the solutions needed to be able to process and analyze data at Grid sites through preparation and submission of jobs to distributed resources and recovery of job output. A standard job submission process performs the necessary environment set-up, executes a CMSSW [43] application on local resources, arranges for any data to be made accessible via Grid data management tools, allows to recover the processing output, and provides logging information for the entire process. Over the past years, the architecture has moved to pilots and CMS relies on the Glide-InWMS for most pre-processing functions. At the application level depending on which is the nature of the processing jobs, different architectural needs arise and CMS implemented two separate systems: WMAgent [44] for centralize production processing, and CRAB for distributed Grid analysis. The description of both fo beyond the scope of this thesis, in particular the former. The latter is quoted in the following, so it is briefly introduced below.

The CRAB [45][46] is a CMS dedicated tool for workflow management for analysis jobs. Its main feature is allowing users to submit jobs to a remote computing element which can access to data previously transferred to a close storage element. Its main functions are: interfacing with the user environment, data-discovery/location services, job execution and monitoring, output recovery and out data transfer to final destination (the Asynchronous StageOut component). Via a simple configuration file, a user can thus

access data available on remote sites as easily as he can access local data. A client-server architecture allows the jobs to be not directly submitted to the Grid byt to a dedicated CRAB server, which, in turn, handles the job on the behalf of the user, interacting with the Grid services. This allows to insulate all retrials and resubmissions, thus simplifying the user experience.

The Data Bookkeeping Service provides access to a catalogue of all event data from Monte Carlo and Detector sources and records the files and their processing derivation. With the information in the catalogue it is possible to track back to the original RAW data or Monte Carlo generation. Data files are mapped to File Blocks that pile related files for data placement purpose, their location is tracked by DBS too. The system is built as a multi-tier web application, and is used for distributed analysis, production data processing activitied, and associated with the data location in PhEDEx. The data managed by DBS are not actually data but pointers to physical data organised by other parts of CMS data management system. Typical uses of DBS include MC generation, detector data, large scale production processing and user data analysis. The data have to be transfed among several service levels, this operation is took over by PhEDEx (Physics Experiment Data Export). This project is a source for large scale data transfers across the Grid. PhEDEx provides a centralized system for making global data movement decisions and a realtime view of the global CMS data transfer state. Many of low level tasks (large-scale data replication, tape migration...) for CMS are automated. PhEDEx is composed of a seried of autonomous, persisten processes, in PhEDEx terms agents. The agents task is to share information about replica and transfer state through a database.

2.3 The CMS dataset popularity

The data management is a very hard challenge, especially when dealing with huge amount of data; one take account of limited storage capacity, not all data can be hold forever. The CMS experiment has collaborations all around the world which submit everyday about 200000 jobs, it entails considerable application of Grid workload and resources to manage corresponding data. Thus, a big goal is to create a compunting model for the optimisation of usage and storage availability through automated procedures. The main purpose is to make a transition from static data placement to dynamic one. CMS developed this project, the CMS Popularity Service, learning from ATLAS similar experience. A useful brand new concept is introduced the "data popularity". A early topic was about the PhEDEx Service [47] whose feature is to create and distribute the files copies at each site; then, many users may keep access to the same files, but for the storage space sake how is it possible to decide which copies are useful and which are not? In this task the "data popularity" concept occurs; it is a misurable parameter able to quantify the interest of the users analytics for data exiting from MC simulation or data samples; monitoring the number of access and upshot to files by users' job. The CMS Popularity

Service tracks the time evolution of : dataset name, number of access, success or failure of data access, CPU hours, number of each users that execute the access. This system is still in progress, currently the informations collected are used to trigger ad-hoc cancellation of least used replicas and trigger ad-hoc replication of which one is considered most popular. A long term goal is to improve this model with the target to be adaptive and able to predict future behaviours of the CMS systems from the monitoration of their performances in the past.

Data treatment

The CMS experiment during both the run1 and run2 has acquired plenty of data; the data used for analysis go under transformation in format and reduction in content (only essential informations are taken). The computing process goes as follow: the collisions data are streamed to HLT (High Level Trigger) and then organized into trigger streams. They are collected at the Tier-0 center and allocated to CMS Analysis facility (CAF) at CERN and Tier-1s. Later, a portion of those data are moved to Tier-2s for simulation process (MC generation). The final step are analysis tasks at Tier-3s. All these data, as above, are replicated in multiple copies (PhEDEx) and accessed by analysis groups using the WLCG services; in CMS is mostly used CRAB. The data are logically organized into run,files, blocks, and dataset. Part of those data (see below) refers to operations themselves, monitoring data, machine logs etc. but they are rarely accessed and analysed, because more focus is given to near-time debugging purposes than a study of time trend. In addition those data result in a dataset that needs a data validation and cleaning process before being suitable.

Type of data

There is another type of data and metadata concerning the performances of the computing operations, this is an heterogeneous ensemble of non-physics data known as *structured* and *unstructured*. In common jargon the structured data refers to information with high degree of organization; in CMS it is a collection of information about CMS Computing activities and are suitable via CMS data service APIs (Application Program Interface). For example the already quoted PhEDEx transfer management database is a source of structured data, and the DBS system (the CMS source for physics meta-data); further examples are: the Popularity Database for dataset user access information; SiteDB collects information about pledges at WLCG sites, deployed resources; the CERN Dashboard that is a big repository of details on Grid jobs etc.. The unstructured data type on the other hand, generaly doesn't have a definite location because cannot be easily stored in a database, in despite of this difficulty usually it is rich in content; unstructured data for example are HyperNews forums, CMS twikies etc. Alongside there is also the *semi structured data* which is a kind of information not located in a relational database but

at the same time it can be analyzed easly, follows a sort of scheme in organization and can be stored in a database; it can be found from CMS web logs, calendar systems.

Relevance and application of data popularity

The data popularity is a metric of interest for efficient data placement strategies based on users' activities. The easiest way to understand end-users' interest is to invesigate the datasets popularity, infact these latter are usually used as principal unit in data analysis process and they are also the final product of this chain. So, can be stated that the knowledge about the nature of CMS dataset popularity aims to optimize the computing model and reduce its operational cost. A question arises as of which is the "best" way to define the popularity: in practice, any definition will work provided that its application allows to produce predictions of practical use. To build a proper definition, the popularity DB provides a quite wide range of metrics: the number of accesses to a given dataset, the numer of users/day recorded in accessing a dataset, the total number of CPU hours spent accessing a given dataset, along with normalized values of those attributes over full number of datasets. Depending on the specific kind of usage of the popularity data, definitions may change and require work to define the most adequate one for the objective of the study.

Chapter 3

Studies of CMS data access patterns

First of all it is usefull to explain how the collected data are arranged and organized, and general feature of the plots we will present and comment.

In this work we use popularity data from 2014 to 2016. For every year, the "data keeping period" (the time of the year in which we collect and keep the data for further analysis) ends in March, June, September and December; so, data is collected and analyzed every quarter.

Subsequentely, data is analyzed into three categories that we call "time windows": 3M, 6M, 12M. These classes refer to 3, 6, and 12 months in the past starting from a given month and year. Finally, data are sorted by number of accesses. As a final outcome, we are interested in the volume of data, both accessed and non-accessed, within each time window in the past starting from the end of a given data keeping period. The goal is to use this general information related to the data volume hosted on Grid storage to study the CMS data accesses and try to extract particular usage patterns.

In the following, several graphs will be shown and explained. The format of some of those plots is the same required to be shown at the WLCG Computing Resources Scrutiny Group, a formal body whose task is to inform the decisions of the Computing Resources Review Board (C-RRB) for the LHC experiments. These plots - that will be quickly referred to as "scrutiny plots" in the following - show the fraction of data volume that was accessed 0 times, 1 time, or more. In particular, the data volume with 0 accesses is divided in two bins, the so-called "0-old" bin and the so-called "0" bin: they all contain volumes that recorded 0 accesses, but the former groups the existing data older than the given time window, while the latter refers to new data produced (thus, newer than each given time window). Their utility will become more clear in the following sections.

3.1 Global view

Given the premises as in the previous section, in this section the focus is the study of the patterns of CMS data accesses with no particular breakdown into any specific data type or any specific WLCH Tier, i.e. we will consider access to CMS data to all data types interesting for analysis on all Tiers where they may reside. A global view is interesting as it allows to get a first overall picture of how the CMS storage is actually used in analysis.

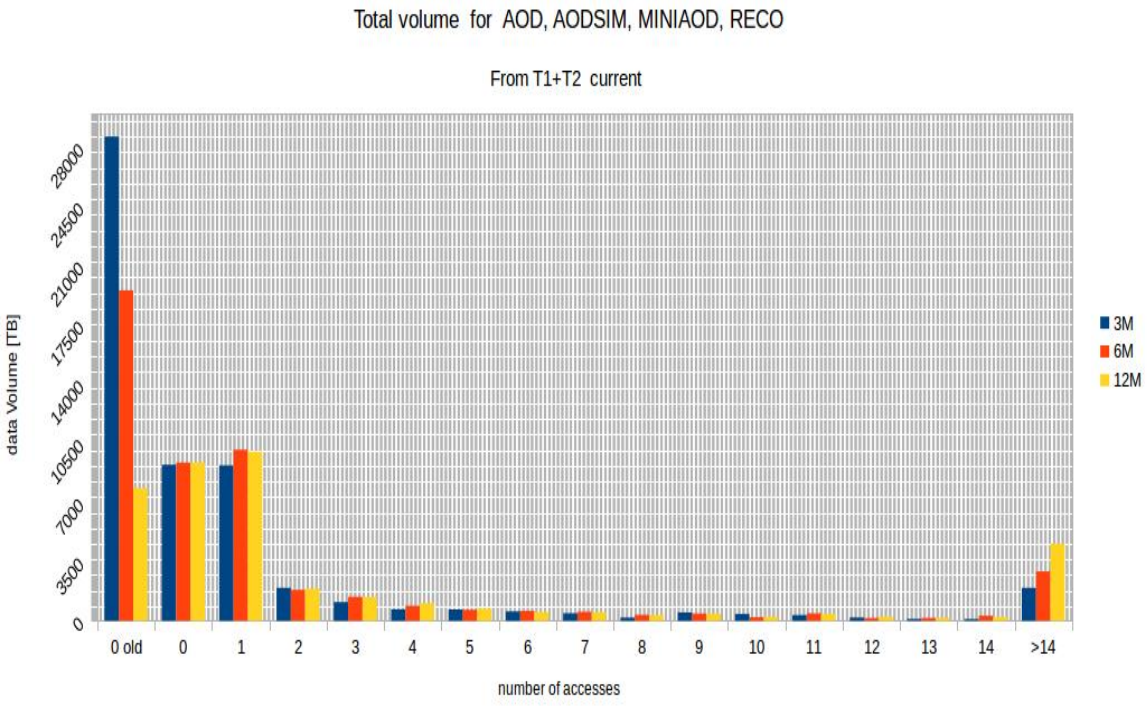

Figure 3.1: The scrutiny plot (see text for explanation) showing the total data volume (AOD, AODSIM, MiniAOD, RECO) accessed from Tier-1 and Tier-2 sites as from September 2016, subdivided by time window and number of accesses.

As initial example we consider the current data keeping period (September 2016). In Figure 3.1 a scrutiny plot is displayed, and it allows to perform a few considerations. First, we can see that the only bins of interest are 0 old, 0, 1 and > 14, the orthers being much lower is size, and more regular, hence less relevant. The yellow bar, which refers to 12 months in the past (from September 2016), provides this information: if we look back from September 2016, we see that there are about 8PB older than 12 months and never used; about 10PB younger than 12 months that in these 12 months have recorded zero

accesses; more than 10PB of data with only 1 access in the last 12 months; the rest is accessed more than 1 time in the last 12 months, peaking at about 4 PB accessed many times (> 14 in the standard scrutiny plot). A quick comparison with the other time windows is instructive. We can see that tightening the time window, the > 14 bin decreases, ad this is explainable because over 12 months there is more probability for this data to have been accessed than there is over only 6 or 3 months in the past. In particular, the 12 months window will tend to be very inclusive (sort of "catch all") whereas the 3 months will tend to be more exclusive (sort of "tooshort time in the oast to reflect actual analysis interest"). This leads us to first preliminary but interesting observation, that we will also apply in the rest of the thesis: among the considered three time windows, the 6 months time windows is the most realistic estimate of what really happes in terms of data access.

Back on Figure 3.1, if we look at the 0-old bin difference between the 12 months time window and the 6 months time window, the latter is twice larger: the more you look into the past, the smaller the 0-old bin will become, but it is still of considerable size. This is an indication that regardless the old data deletions from disks that CMS performs, there is still sizeable fraction of "old data" that remains on disk and is unaccessed. This information has a value when analyzed in terms of its dependence with time (see in the following sections).

Back on Figure 3.1, if we look at the 1-bin, we observe that its size in the 3, 6, 12 months time window is roughly 10PB (little less in 3M). A first observation is that being this bin much higher than most other bins apart from the 0-bins, it shows that CMS analysis teams tends to have on average just one submission pass over interesting data, from which they produce derived data which they further analyse (and those accesses are not counted/showed in this kind of plot). A second observation is that the size of the 1 bin and the 0-new bin is comparable in each time window, implying that it may be concluded that the fraction of data younger than any selected time window have a 50% chance of being accessed once or a 50% chance of being left unaccessed in that time window.

At this point, it can be instructive to deviate from the standard scrutiny plot format and face the problem in a different way: display, in different time windows in the past from a given time (a moving reference, of course), only the datasets that have been accessed at least once (i.e. > 1 bins altogether) and the datasets that have never been accessed (i.e. 0-old and 0 bins altogether). One example is shown in Figure 3.2 where the observables above are shown for accesses in the last 3, 6, 12 months and at all Tier levels of interest (Tier-1 and Tier-2) for all data types of interest, i.e. AOD, AODSIM, MiniAOD and RECO. This allows to perform a few interesting considerations.

First of all, the sum of non-accessed data volume and accessed data volume in each

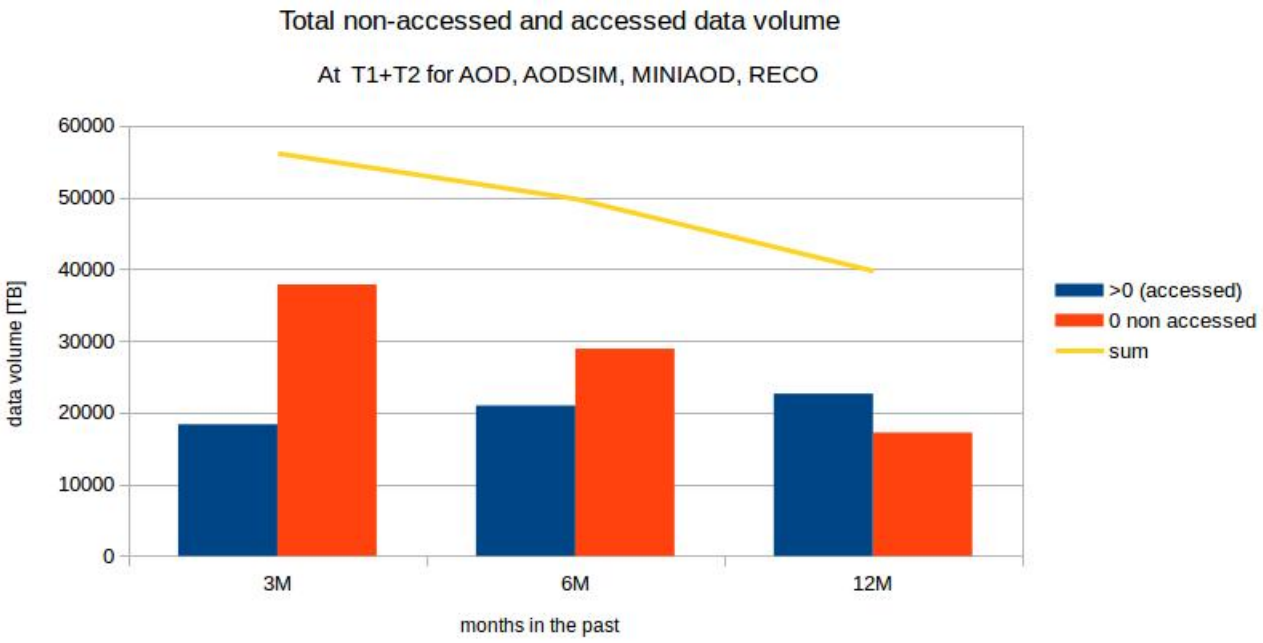

Figure 3.2: Same data as from Fig. 3.1, but displayed with aggregate bins (see text for explanation). The yellow line displays the sum of the red and blue bins, which show the non-accessed and accessed data volumes respectively, for each of the three considered time windows in the past, starting from September 2016.

time window should correspond to the total existing data volume on disk in that time window - as there is no data that can be in a condition that is not one of these two mutually exclusive categories. While is true inside a specific time window, this may well change from one time window to another: the total data volume is impacted by data proliferation as well as data deletion, and both happen at different time windows in a largely unpredictable manner (i.e. in ways related to the CMS experiment activities and needs). A pragmatic approach to evaluate how the situation is evolving could be to measure the total data volume in each time window (the yellow line in Figure 3.2) and see how it evolves over time (i.e. from the 12M time window, through the intermediate 6M time window, to the 3M time window - note, going from the right-hand side to the left-hand side in Figure 3.2): if the yellow line shows an increasing trend, it means that over time data was created more than deleted, whereas if it decreases, data was deleted more than created. This information is useful because then one can check the fraction of accessed data volume, knowing the trend in the total existing data volume, and thus being in a condition to draw some conclusions.

To explore this, more insight is needed and we must go beyond the general overview across all data types and WLCG Tiers, which was the focus of this section. An analysis of the patterns we observe can hence be attempted by considering that the popularity data has been collected for various data types (AOD, AODSIM, MINIAOD, RECO) and for accesses at Tier-1 sites and Tier-2 sites separately. We are in good position then to explore data access patterns we observe, according to different views, and this is explored in the next sections.

3.2 Evolution with time

In this section, the focus is the study of the patterns of CMS data accesses with no particular breakdown into data type or WLCG Tiers, i.e. the "global view" as from the previous section, but studying how it evolves over time. This is done by studying sliding time windows, i.e. 3, 6, 12 months in the past starting from the present, from 3 months ago, from 6 months ago, etc and going 2 years into the past. The key point of the study of time evolution aspects is that we can envision the whole set of data as a flux and we are interested in its ensemble behavior, namely we would like to verify is observed patterns would stay unvaried or not, and what this would imply.

Unlike the plots in the previous section Figure 3.3 represents the evolution with time of the global view: now we focus on how bins contents (that is a ensemble of AOD, AODSIM, MiniAOD, RECO data tiers) is evolving per data keeping period, highlighting the time windows of 3, 6, 12 months in the past. Despite the plot contains plenty of details and it is hard to read, global behaviours can be extracted from it and they help

in assessing how the storage management health is currently in CMS. All the bins as from the previous section are displayed in different colors, but we discard all bins from 2 accesses to 14 accesses, as their flatness relveals a quite regular activity which is not particularly relevant in the current study. We emphasise only the following bins: 0-old, 0, 1 and > 14. Let's re-state the importance of each bin in terms of what it is intended to display, so we may settles on what an "expected" behaviour over time would be, and then check with the data if we indeed observe this expected behaviour or deviations from it.

The 0 old bin, for the reasons we already stated in the previous section, can be assumed as a parameter that tells how good CMS is in cleaning-up relatively old data that still occupy disk space (e.g. LHC Run-1 data fall for sure in this category). The 0 bin can be conceived as a parameter that tells how good CMS is in making sure that the data collected and produced relatively recently are indeed accessed from disks (e.g. data just produced may well not be accessed yet, but data produced some time ago should have been recorded some accesses). Quite another this is the bin 1, as it represents the data volume in each given time window with just one access (and in many case the number 1, as discussed earlier, is not representative of the real activity over the dataset). In an ideal world a perfect placement strategy must aim to avoid the 0-old bin (i.e "keep disks clean from old unaccessed data"), to reduce the 0 bin to a decently low level, and to increase ad needed the size of the bins from 1 to N. This situation, if observed in this kind of plots, would reflect a more efficient use of the Grid storage, even in the constant presence of newer data being recorded and accessed.

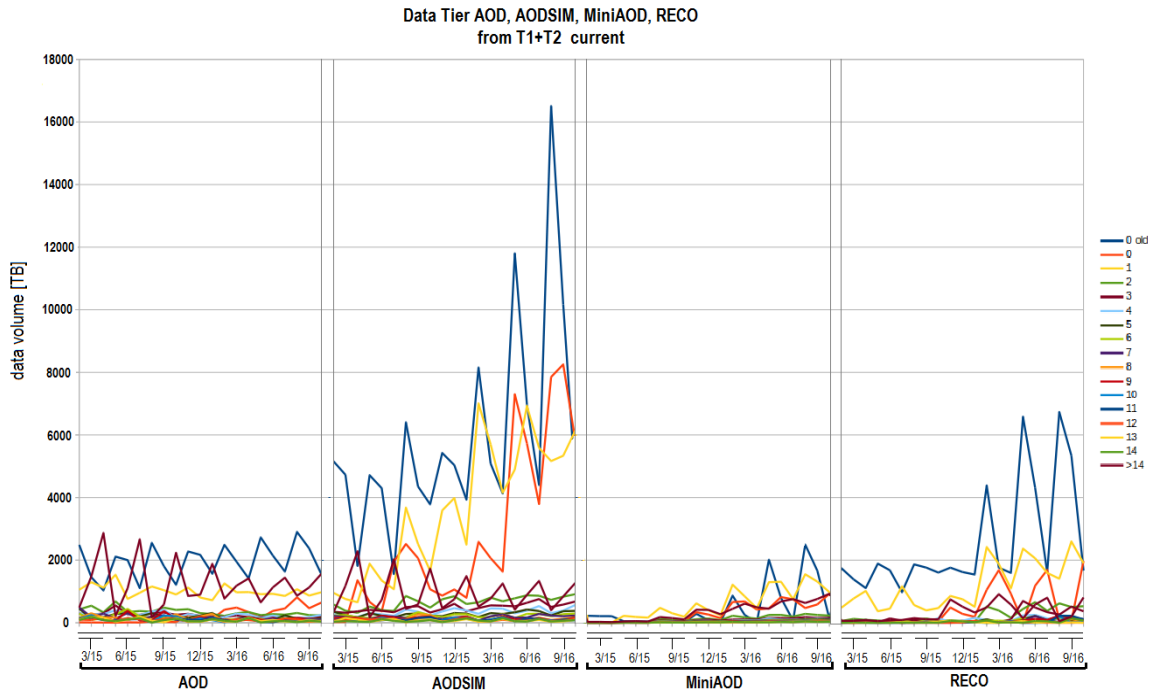

Figure 3.3: Trend of number of accesses per data keeping periods, time windows and data tier.

With these premises, from Figure 3.3 we can speculate about some aspects:

1. A first consideration concerns the AOD data type. In general looking inside each time window in the past from a given period we can state that the oscillation between non-accessed and accessed data volume are well balanced. Looking through the whole set we can notice that bin > 14 is decreasing and bin 0-old is increasing between June and September 2016. In fact at least 75% of analysis operations can be performed on MiniAOD but there is still a small part 25% of the analyses that have not migrated and still use AOD. We know that the MiniAOD format has been introduced in CMS in late 2014, and indeed one sees aforementioned effects of their presence in late 2015: for AOD during March-December 2015 there are more accessed data than non-accessed ones. Moreover, in this same period, the orange line (that rapresents the 0 bin) is quite flat: this means that most recent data used to be accessed relatively quickly in unpredictable distributed analysis; then, in late 2015 the behaviour slowly changes and the accesses started to be apparently more slowly. A first qualitative example, the 0-old and also the 0 started to grow more slowly. A second quantitative example, considering the 12 months time window in March 2015 and the 3 months time window in September 2016, in two years, the data volume of the > 14bin is decreased of 69%. These observations on AOD will be confirmed by the observations on MiniAOD (see bullet 3 below).

2. A second, indipendent consideration is about the situation of AODSIM. The figure shos that this is evidently becoming more and more critical over time. The bins 0-old, 0 and 1 are increasing considerably over time, while the > 14 bin is constant. The AODSIM data tier has simulated event content produced by Monte Carlo simulation, and it is known that in CMS lately its amount (and access) has increased. In this case the 0-old bin exceeds 16PB during the last 3 months since September 2016, but such data volume is consistent with the fact that there has been an ingent production activity as can also be inferred by the 0 bin increasing trend (i.e. data produced recently and not yet accessed). However, as shown by the plot, the overall CMS space management has gradually worsened (e.g.see the situation during the 3 month in the past from March 2015 when the non-accessed data volume was already about 5PB despite of 0 bin and 1 bin which were about 0.5PB and 1PB respectively). We also notice that the 1 bin has (in the case of AODSIM) more entries than other bins in many time windows over the last 2 years: it may look quite odd that data of interest are really used just one time, but this may just be an effect of the fact that this data is used as initial step by analysts and then they use only the outcome, in agreement with what stated in the previous section.

3. The MiniAOD data format, introduced since late 2014, shows a very interesting evolutive path: in March 2015 there were mostly 0 old and then from December

2015 till now we can see an evident ramp-up. This confirms what stated in the first point of the AOD data type. This data tier is growing steadly in terms of accessed since its birth, and in a quite healthy manner: of course there are older data on disk being accessed less, but the volume of data accessed is always larger than the volume of data non-accessed. In a general trend, it is quite easy to note that MiniAOD have reached during Semptember 2016 a volume comparable to AODs. One should also remember that the MiniAOD size is a factor 10x smaller than the AOD, so e.g. a MiniAOD data file mistakenly left (not cancelled) on disk is "wasting" 10 times less space than a similar action done for AODs. Additionally, the MiniAOD data type in this plot must be taked ad a MiniAOD*, i.e. it groups MiniAOD and MiniAODSIM data format, so in principe we should be comparing MiniAOD here with AOD and AODSIM summed up. Now it is clear th impact of MiniAOD on the CMS storage worldwide: the CMS analysis teams (at leats at the 75% level) can perform their usual analysis routine with much less storage space needed, and growing on the Grid in a much more healthy manner.

4. On a last point, the RECO data format is also displayed. This data type is crucial for some dedicated studies but it is of limited relevance to the CMS Grid analysis community, getting decommissioned soon (i.e. same content in AOD and in most cases also in MiniAOD). As displayes in Figure 4.7, early in 2016 the volume of old non-accessed data (0-old bin) started to raise much more than recent non-accessed data (0 bin): RECOs were indeed prodced less, and most of the old ones were kept of disk for last access needs before speeding up with the decommissioning. Becoming soon obsolete, this format is less relevant in this thesis, despite being included as this format is still one of those accessed by CMS analysis jobs on the Grid.

From the observation so far, the role of the information displayed in the 0-old bin seem one of the most relevant. Most (almost all) of the conclusions in this sections can be inferred by looking at the 0-old bin only (e.g. looking at the dark blue line in Figure 3.3). For this reason, it is interesting to transform the visualization as in Figure 3.3 into an evolution with time of the 0-old bin only, and display it in a different way. In Figure 3.4, for each type, each coloured line shows the situation at a specific moment in time (and the 3 points on each line show how this situation changes if one looks 3, 6 , 12 months in the past): in this way, one can observe how the situation indeed changed over time by looking at lines of different colour, and it is easier to catch a peculiar pattern in data and found evidences of what stated above.

Firstly, one can start from the AOD data type: the trend through the years is always monotone increasing going from 12 months in the past to 3 months in the past. In principle, we have also evidence of regular accesses through the year (e.g. correlation with

conferences and holidays), and we could draw more hypothesis about working periods, etc.

Secondly, one can look at the AODSIM data type. By taking e.g. each 3M time windows, a worsening is seen: obviously in this plot the bin 0-new and 1 are not shown by choice, but the behaviour over time of the 0-old alone well supports the hypothesis of CMS space management getting progressively worse as and effect of the presence of this (AODSIM) data tier.

Thirdly, the MiniAOD trend - at least in the periods when it existed already - can be seen indeed ad the motst healthy: no inflation of the 0-old bin is observed over recent times, despite a considerable grow in the MiniAOD data volume over last couple of years.

For the reasons outlined in a previous paragraph, we do not further investigate the details of the RECO access patterns.

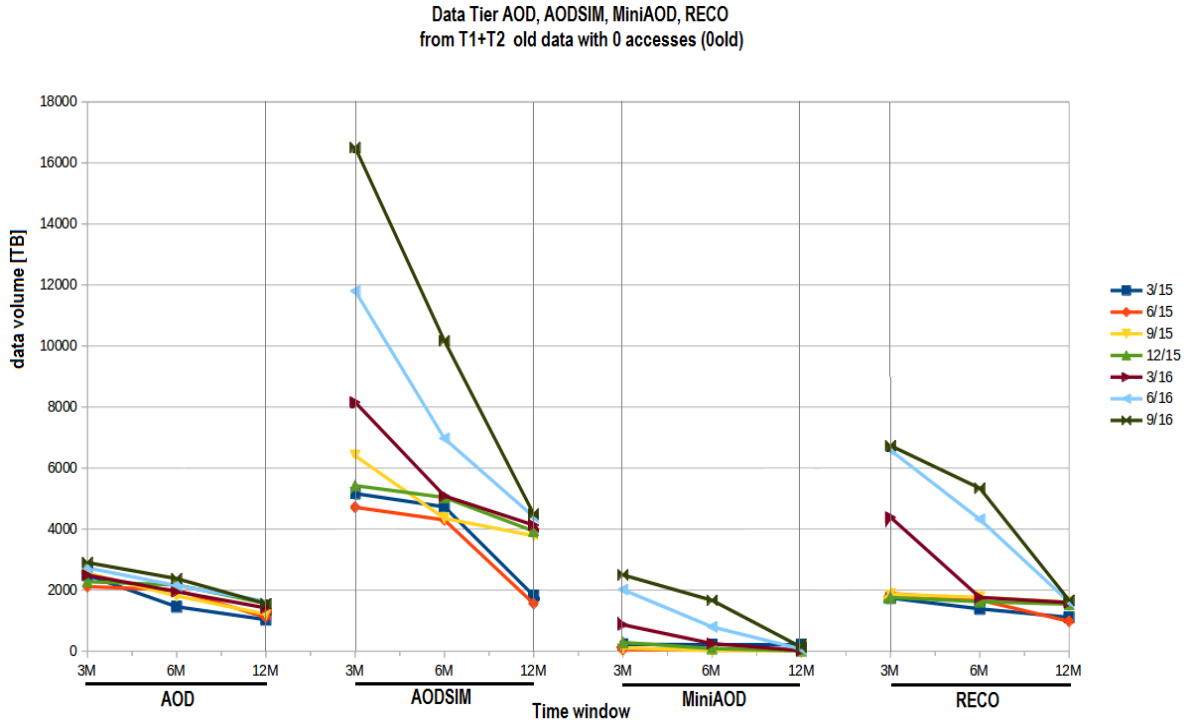

Figure 3.4: The 0 old bin content in each time window for each data tier.

3.3 Data types view: RECO, AOD, AODSIM, MiniAOD

In this section, the focus is the study of the patterns of CMS data accessed with breakdown into the different analysis data tiers: RECO, AOD, AODSIM, MiniAOD. We start from a general study of how and how much a single data tier is used; then we consider different combinations of main parameters and their outcome; finally some considerations deriving from the study of time window. Again, the data used are from both Tier-1 and Tier-2, with no distinction between them.

As a reminder, each step in the simulation and recostruction chain gives information about events which are stored into what we call a "data tier". A data tier may be composed by multiple data formats, then a dataset may consist of multiple data tiers. In the following some considerations about different aspects of each data tier are discussed. One of the reasons for this study is to properly disentangle e.g. the contribution of AOD and AODSIM separately.

First of all, we can make an a-priori consideration about time windows: ideally, data access patterns should reflect the activity of CMS analysts, which on a global scale does not react so quickly after data is available, because realistically few weeks are needed for changes to be reflected on CMS-wide average behaviours. This means that, translating it in terms of time windows, a 3 months window might be simple too short (we got a superposition between non-accessed new and old data); a 12 months window might be too large (it is easy to find more data accessed during long periods); a 6 months window is a good compromise. Moreover it is on average the time window between two major series of conferences, speaking of "Winter" conferences and "Summer" conferences. This latter point will be clarified in the next section. At the time being we consider a 6 months time window as the most significant for the considerations in this section.

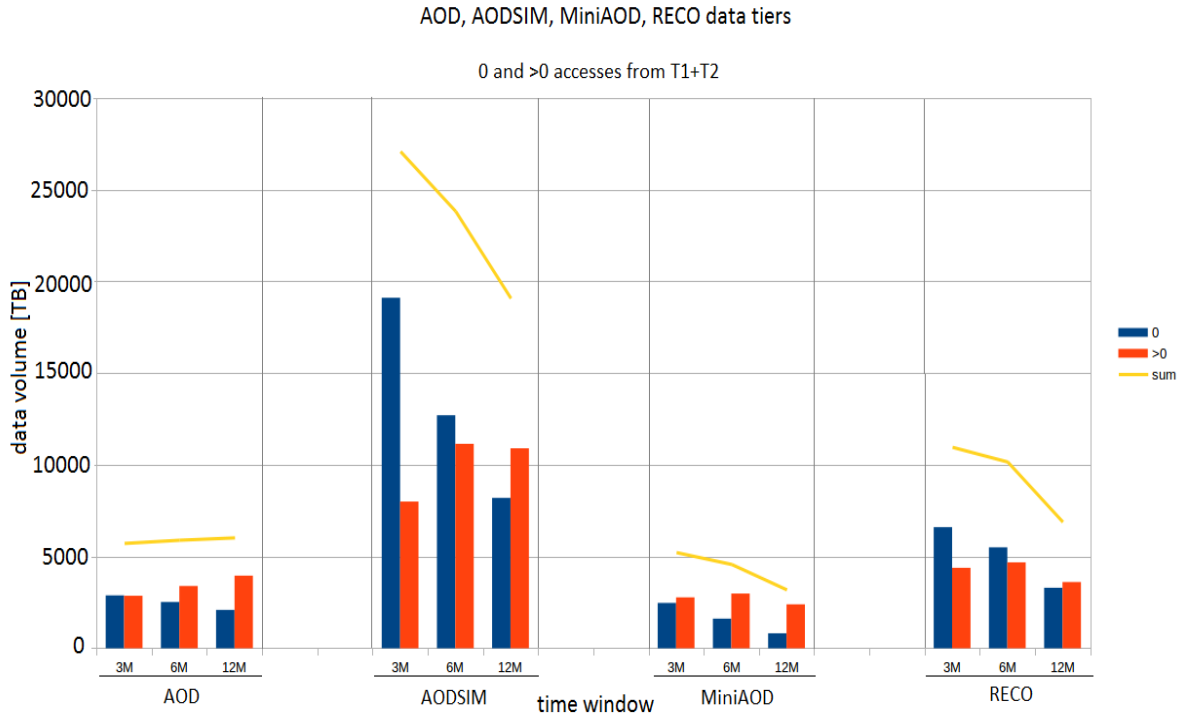

Figure 3.5: Data volume accessed in 3M, 6M, 12M time windows in the past starting from June 2016. See text for detailed explanations.

In the following paragraphs, we will focus on the one data tier at a time, and discuss it throughly.

AOD

As discussed previous, and also shown in Figure 3.5, the regularity of AOD total volume (i.e. sum of 0 accesses and > 0 accesses volumes) in a fixed period is evident: there is a decrement of 0.005% in total volume going from 12 months to 3 months in the past, hence totally negligible. The previous consideration about the adequateness of the 6 months time window applied to the AOD case: as we can see in Figure 3.5, during this period the analysts have accessed, over a total of roughly 6PB, about 60% of the data volume and only about 40% has remained non-accessed (whereas the 12 months window would be too optimistic i.e. 2/3 and 1/3, and the 3 months window would not yield yet any useful interpretation of the data).

On Figure 3.6 we describe the 6 months behaviours of the AOD data tier in its time evolution. More (i.e. also 3M and 12M) can be found in Appendix C, anyway in general they are consistent in showing the time evolution pattern of total accessed and non-accessed data volume for AODs: over time (x-axis), looking at 6 months in the past (as

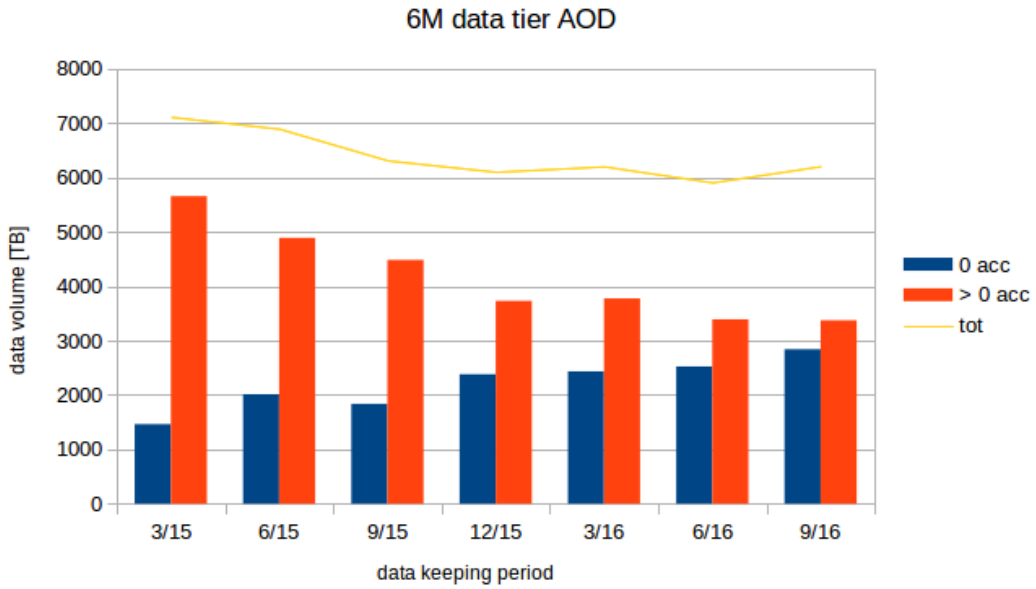

Figure 3.6: Data volume accessed vs non-accessed in the 6M time windows starting from different data keeping periods, for the AOD data type.

in Figure 3.6) one evidently sees that AOD data is accessed less until it reaches a plateau (since late 2015), while the non-accessed volume increases still (see also Section 3.1.1). This implies that most recent data tend to accumulate on CMS Grid disks and are being accessed less and less over time. This is an important piece of information because the AOD tier is very important (it contains a copy of all high-level physics objects and a summary of RECO information) and it has been the dasis for CMS distributed anlysis for long time, but something is happening over last couple of years. The total volume is anyway always - on average - between 6 PB and 7 PB: the total size decreased over the years, but so slightly that it can well be considered basically flat. A primary feature of the AOD access pattern is hence its small fluctuations over time.

Additional considerations can be made by looking at the 0-old bin (Appendix A) and 0 bin and 1 bin (bot Appendix B). E.g. by comparing Figures A.1 and B.1 (in Appendix A and B respectively) we note that the volume of the 0-old bin is never under 1 PB whereas the 0 bin neve exceeds the 0.8 PB. Thus our assumption about a quick creation and utilisation is correct, so the problem is about the 0-old bin increasing. Then, looking at Figure B.5 (in Appendix B) one can use it as evidence that the 1 access bin gives a large contribution to the overal accessed data volume, and this increases the interst towards bin 1.

Digging even deeper, one can look at the figures referred to AOD 0-old, 0 and 1 (in Appendix D and E respectively) for a single time window evolving with time. There is no evidence of seasonally in bin 0 old and 1 (seasonality would indicate that a schematic list of activities tha occur every Δt happens in the same way and just repeat in cycles). Evidence of lack of seasonality observable in the data is shown in Figures D.1, D.2 and D.3.

AODSIM

In Figure 3.7 we clearly observe for AODSIM a completely different trend with respect to AOD. First of all, we have seen already that from the 12M time window to the 3M time window the AODSIM data volume on average increases, so data are created more than deleted; but Figure 3.7 shows that at the same time, over the years the fraction of non-accessed data is steadily increasing - especially in 2016 - while the fraction of accessed data is sort of plateauing, if not even decreasing just recently. Additionally, the speet at which the non-accessed data volume increases is mush higher than any speed of any change in the accessed data volume. This means that the data proliferation of AODSIM at T1+T2 is such that over time this data type is populating more and more the disk storage at sites while not being accessed so much: over last 12 months, 8PB of data are shown to be non-accessed, and the same number increase to become almost 19PB over just last 3 months. In this case, not only Figure 3.7 but also Figure C.7-9 in Appendix C are consistent in highlight this as a worsening process in CMS. Note anyway that the amount of AODSIM in the CMS storage systems is heavily increasing over time - it is a huge difference with respect to AOD, which were at a roughly flat level.

From former considerations, we can point out that the pattern of non-accessed AODSIM data volume is the most interesting, so we can elaborate more exploiting additional plots which can be found in the Appendix.

Firstly, we can study how the 0-old bin has changed (in time windows of 3, 6 and 12 months) since early 2015 till now, see e.g. Figure A.2 in Appendix A. One first observation is that the lines corresponding to June and September 2016 show a net detachment from earlier periods, in the fact that the total volume passes from roughly 8 PB in March 2016 to 12 PB in June 2016 and to more than 16 PB in September 2016: 4 PB (8 PB) are a significant variation in 3 (6) months only. Moreover one needs to add the new non-accessed data that, during September 2016, has reached 8 PB. As discussed earlier, it is not so odd to see a gap between 2015 and 2016, due e.g. to the MiniAOD introduction. In this case, it can again be seen that the 1 bin is comparable with the 0 one but it is also interesting to see the difference between the lines representing March, June and September 2016. In this situation we see that during the last 6 months from September and June 2016 the volume of data accessed just one time is about 6PB instead of 7 PB during the last 6 months from March 2016.

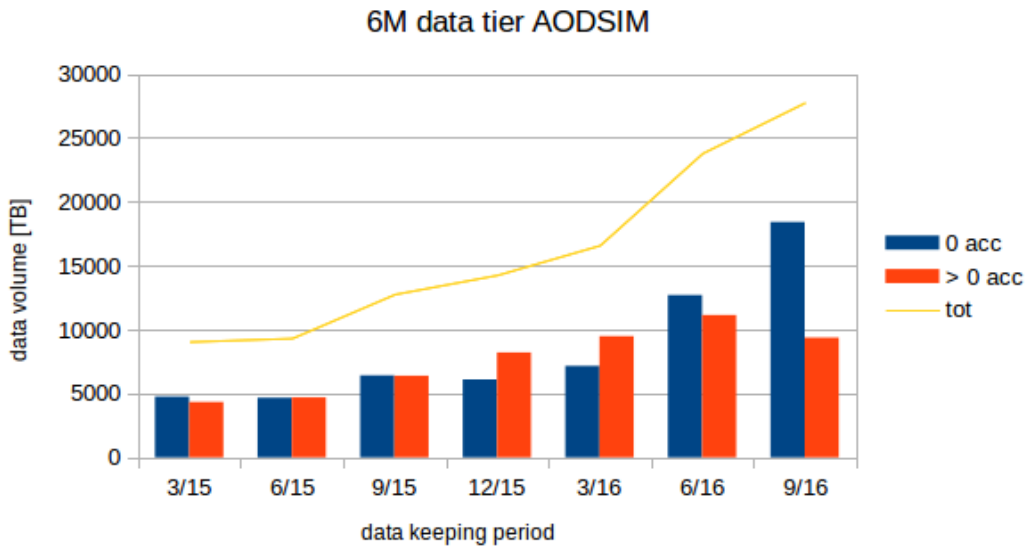

Figure 3.7: Data volume accessed vs non-accessed in the 6M time windows starting from different data keeping periods, for the AODSIM data type.

Another very interesting observations on the AODSIM trends can be done by selecting 3M, 6M and 12M by themeselves and display their trend over the entire period under study, as in Figure D.4, D.5, D.6 in Appendix D. It can be seen that for both 3M and 6M the value of volume is quite regular and semi-flat until December 2015, when plots start to show a monotone growth; of course correlation between 3M and 6M would be expected, because there is a superposition between the two data samples, but the trend visible in either of the two plots is a convincing evidence in itself of a "change of place" in the quantity of non-accessed AODSIM data volume, and these plots stand as a portrait of a critical situation.

MiniAOD

An important feature of the MiniAOD format, as stated previously already, is their smaller size (10x less with respect to AOD). And a remarkable feature of their observed access pattern is that the volume of MiniAOD accessed data always exceeds the volume of non-accessed data (except for few months just after their introduction in CMS). From Figure 3.5 it was observed that the total MiniAOD data volume on average increases over time; the largest contribution to this comes from the accessed MiniAOD. In early 2016, looking in the previous 6 months, more than 2/3 of the existing MiniAOD were accessed at least once; over time this ratio slightly reduced of course, as more MiniAOD were produced (and more versions of the same physics content but with fixes and updates, thus driving users to access only last version), but it is still remarkably high

today. In a nutshell, this is an agile and useful data format: MiniAOD are generated quickly, and quickly used. Today, the MiniAOD format is almost at the same level (in terms of global size) of the AOD format and moreover its utilization in analysis seems to promise CMS a by-far more performant and better-manageable system in the long-term.

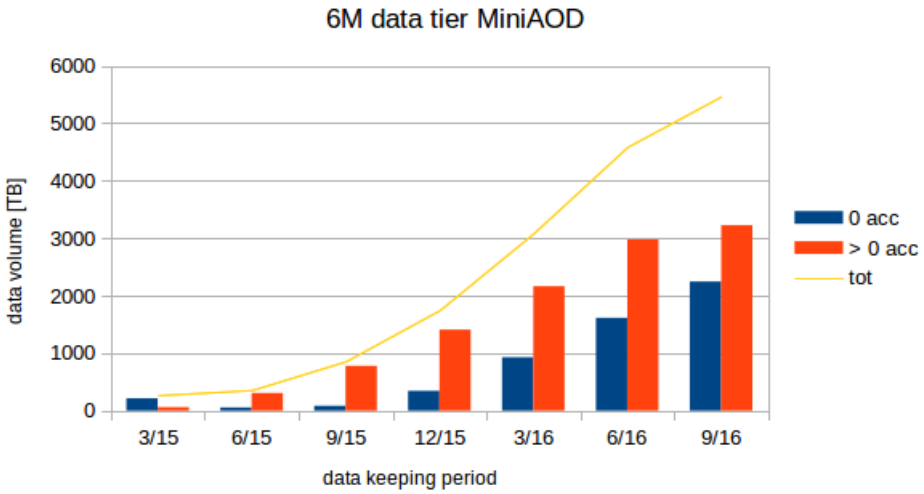

Figure 3.8: Data volume accessed vs non-accessed in the 6M time windows starting from different data keeping periods, for the MiniAOD data type.

For additional insight into MiniAOD, we can refer to Figures D.7, D.8, D.9 in Appendix D, that show separately the 3M, 6M and 12M time windows in the past for the 0-old bin. The trend displayed is really interesting: it is evident that the MiniAOD data volume evolves over time from June 2015 to March 2016 in very similar ways in the 3M, 6M and 12M time windows. Another remarcable thing is that if we select the 3M, 6M and 12M for same data keeping period March 2015 the value of volume fluctuates around the same value 200 TB. The Figure 3.9 shows not only the singular time window but also the subdivision per number of accesses; Figure 3.9 is referring to current period but it is well representative of the MiniAOD general trend. Over 6 months in the past with respect to September 2016 the data volume accessed by analysts more than 14 times is 0.8 PB.

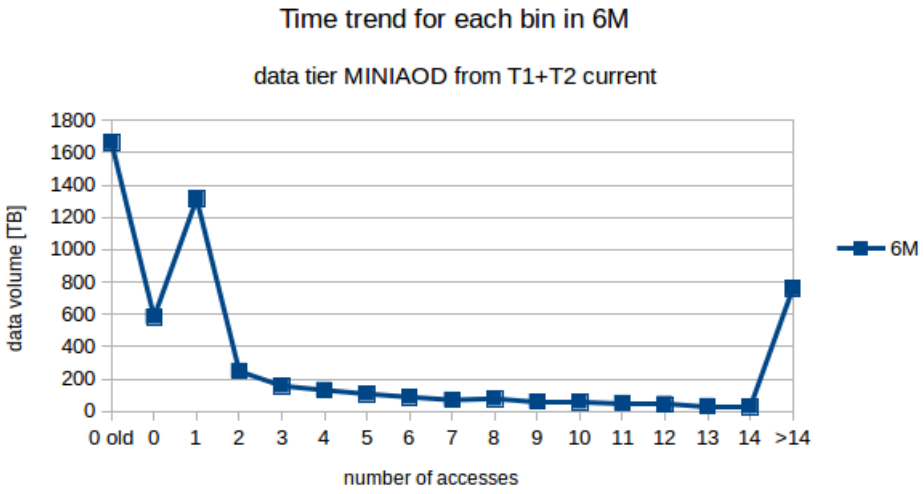

Figure 3.9: General trend (per number of accesses) for MiniAOD in 6M time window from September 2016.

RECO

Any insight on the RECO data type should not take long at this stage, as it has been discussed earlier that this format is going to become transient soon, and not written to disk anymore. Few observations follow, anyway.

Because of this, we would expect to see: an increment in the 0-old data volume (because this format will soon be decommissioned and older non-accessed RECO could pile-up); a small 0 accesses bin (because we produce less RECO, so there is more focussed access to fewer data); a small increment for bins with more than zero accesses (because less and less analysts will access this format). We can refer to Figure A.4 in Appendix A to check these a-priori expectations. We note that the 0-old bin has indeed the pedicted behavior: in the 6 months time window from September 2016 backwards the volume is roughly 5.5PB and in the 3 months time window it reaches more than 6.5PB. We can see this gap already from the 3 months time window of March 2016 (also the assumption that 0-old data pile-up is right, as can be checked by looking at the step of a bit more than 1PB for the 2014/2015 period). The 0 bin, looking at the 6 months time window (the most realistic one) we can observe that during 6 months in the past from September 2016 the volume of newer non-accessed data is 90TB that is a good correlation with older non-accessed data (we can remark that the 0-old bin content, for the same period and the same time window in the past, and 0 bin are not comparable). All plots seem to confirm the expectations.

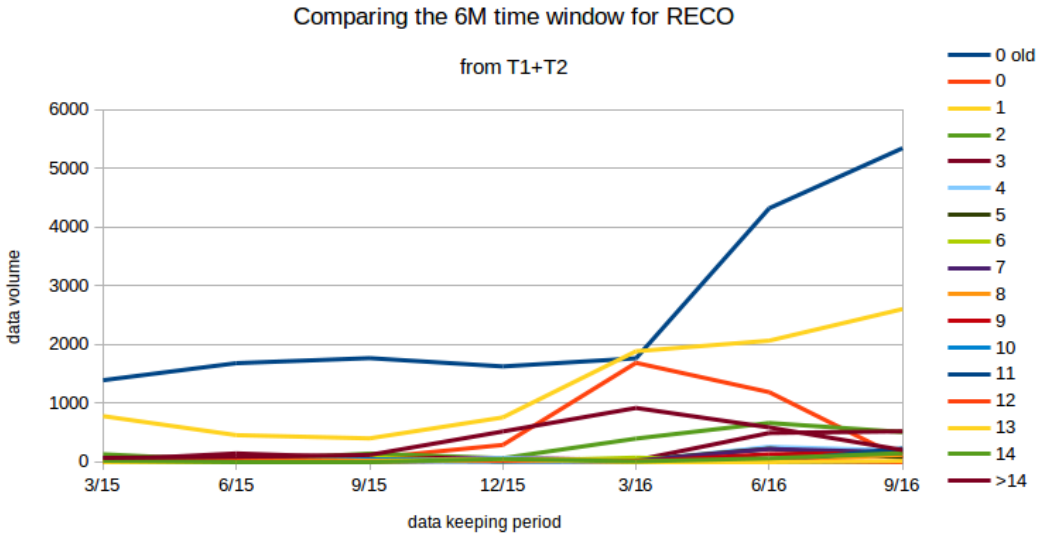

Figure 3.10: Comparison through data keeping periods between bins with different number of accesses.

3.4 WLCG Tier view: Tier-1 and Tier-2

In this section, the focus is the study of the patterns of CMS data accesses with breakdown on the computing sites where the data resides: Tier-1 and Tier-2 (T1 and T2 for simplicity in the following). This is the only parameter left to explore among those considered at the beginning of this chapter: all data accesses have been so far investigated without making any distinction as of whether the data resided at the T1 level or T2 level. This exploration is finally done in this section.

The T1 and T2 roles and activities as from the CMS Computing Model are different (as explained in a previous chapter). In terms of distributed analysis - which is the main focus in this thesis - both T1 and T2 can nowadays perform it, despite the T2 level is the one mostly oriented to this task. The CMS T2 CPU worldwide are on average used 50% for Monte Carlo production and 50% for analysis, with large deviations from the average depending on needs and specific periods. So, in summary, one can assume that the Tier-2 centers in CMS are basically the only locations - besides the specialized analysis facility at CERN - that are designed to host user activities where CRAB jobs access CMS data samples. Furthermore, the T2s are specified with data export and network capacity to allow such centers to refresh the data disk storage quite regularly for analysis, upon needs.

Starting from an overview of how the activities are distributed, a look at the LHC Sched-

ule e.g. in 2015 or 2016 could give some insight as of when the analysis may happen on the basis of when LHC data is being taken and recorded by CMS. But attempting to make this correlation is most probably failure prone, as the distributed analysis - despite admittedly peaking in some periods - are in general relatively active throughout each year. And the periods in which it might actually peaks are only somehow predictable as a function of the international conference to which the experiments communities aim at presenting results, more than the data taking periods themselves.

It is hence more cautious to explore a set of parameters we know from previous considerations, and try to disentangle the access patterns as a function of the WLCG Tier level in which they happened. Some discussions along these lines follows.

In order to investigate how the differences between T1 and T2 sites reflect on data access patterns, we followed this approach. Firstly, we directly compare T1 and T2 accessed and non- accessed data volume in different time windows for each data tier. Then, Now, we start with an insight of each data tier into the two tier level; more attention is given to how the total volume changes, and to which tier level (according to data) gives major contribution to the total pattern (Tier-1 + Tier-2) studied in the section before.

AOD

A first thing to check for the AOD data tier is the total data volume accessed and not-accessed at different Tier levels; then, for accessed data we highlight the bin with 1 and > 14 entries that in previous descriptions showed some interesting aspects. For most figures, we refer to Appendixes. By "global or "T1+T2 in the following we refer to aspects we observed in the totality of Tier-1 and Tier-2 sites in previous sections/chapters.

Firstly, we check the T2 level and investigate AODs at T2 with respect to T1+T2. From the Figure 3.11 and F.1, F.3 in Appendix F, we can note that the regularity of AOD format at the T2 level is preserved as it was at the T1+T2 level. As usual, 6 months is a stable situation to check, and indeed the variation in number of accesses is such that the total volume stands in the interval [5 PB, 6 PB]. If we assume the 6M time window as a reference, Appendix H Figure H.1 shows that at the T2 level the use of the AOD data format is following an apparently predictable trend (Appendix H, figure H.1). Note that data for September 2016 are missing yet for the T2 level, and this caused the drop in the plot, which has no other meaning than this.

Secondly, we compare the T2 level with the T1 level for AOD, shown in Appendix G, Figures G.1, G.2, G.3. Looking at the very same plots for the Tier-1 we note that the total data volume trend versus time is not flat as for T2s but it is decreasing in volume for T1s. Additionally, we note that the non-accessed data volume at T1s is a

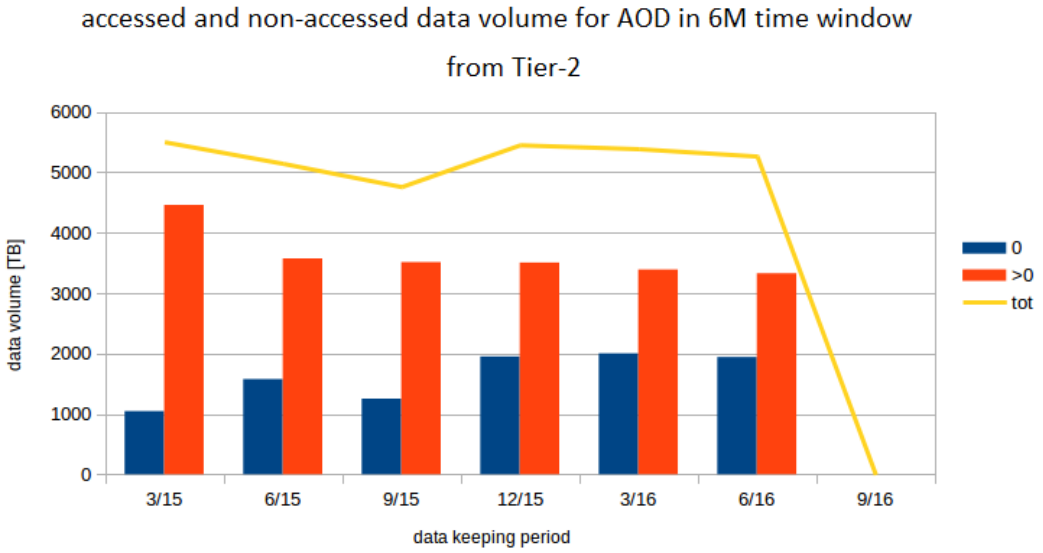

Figure 3.11: Data volume accessed vs non-accessed in the 6M time windows starting from different data keeping periods, for the AOD data type.

big fraction of the total, while for T2s the accessed data dominates. The situation from 3M time window to 6M is quite the same, in particular for recent periods; in the 12M window it changes a bit but we know 12M is not a good representation of the real activity of interest. This is consistent with what we would expect from the CMS Computing model: the AOD data are in multiple replicas in all sites, but it is not uncommon that can be more easily accessed in the Tier level that is ad-hoc dimensioned and prepared for distributed analysis, i.e. the T2 level. So, the total AOD space in CMS is better used when the AOD reside at the T2 level with respect to when they reside at the T1 level.

AODSIM

Firstly, we check the T2 level and investigate AODSIMs at T2 with respect to T1+T2. Appendix F, Figures F.4, F.5, F.6 show a clear evidence of the gradual worsening we also commented globally at the T1+T2 level. If we take for example the 3 months time window (Figure F.4), the non-accessed data volume keeps growing till 2/3 (20 PB) of the total amount for June 2016; the situation is a little better in the 6 months time window, our suggested reference. Thus, we are justified to conclude that the AODSIM non-access situation at the T2 level is currently caused by a bulk of almost 20 PB of data that reside on disks without being really accessed over last few months. To this description we can add the focus of accessed data volume evolution with time respect the 6 months window (Figure.3.12): between March and June 2016 the total number of

accessed data is constant (blue line) at 14PB, also the 1 access bin presence is evident (10PB over 14PB for March 2016) and the > 14 accesses bin is quite constant through the different periods.

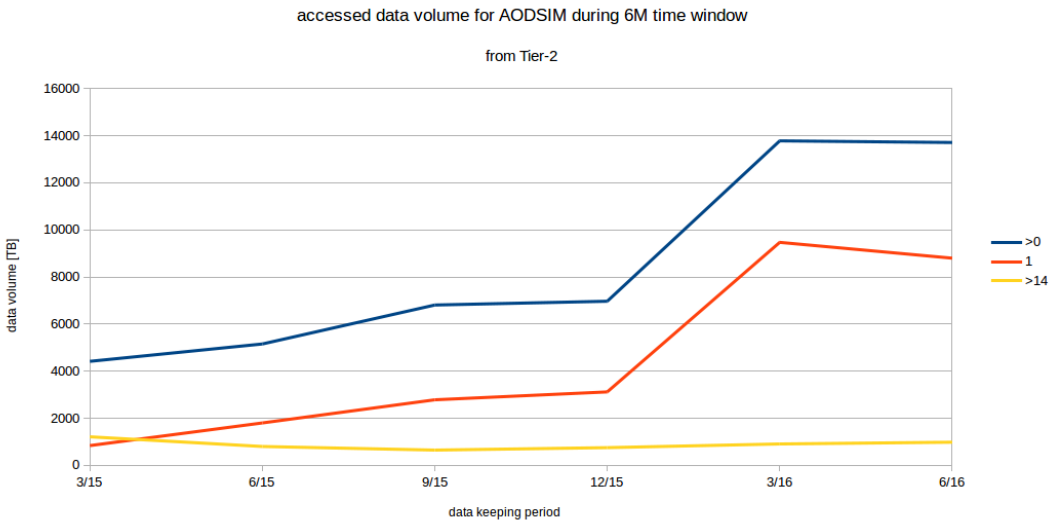

Figure 3.12: Evolution with time of accessed data volume.

Secondly, we compare the T2 level with the T1 level for AODSIMs, as in Appendix G, (Figures G.4, G.6): in each 6 months window (Figure 3.13) the data total volume reaches its maximum then decreases slightly between September 2015 and December 2015, then raises up again. Another interesting fact is that T1s host less than half the AODSIM volume hosted by T2s, and currently - especially in the 3 months and 6 months time windows - this volume in large part is non-accessed. Again we can see this during June and September 2016: the non- accessed data are almost 6 PB, and in 12 months still 6 PB. Now considering the evolution with time of accessed data volume (Appendix H, Figure H.4), we can see that this plot is similar across T1s and T2s. We note also a common characteristic of T1 and T2 regarding the regularity of > 14 accesses bin (of course they are different in volume).

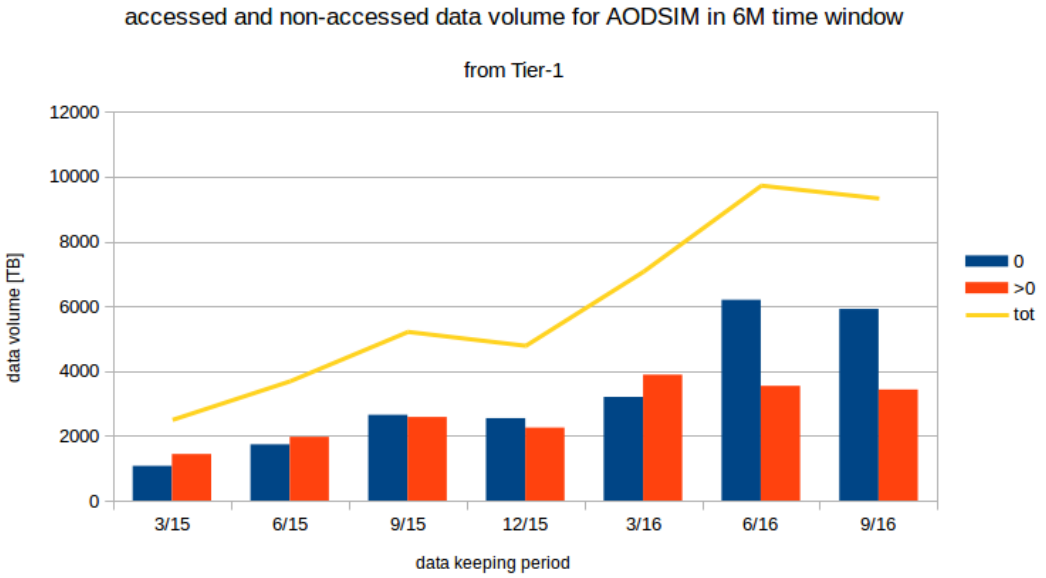

Figure 3.13: Data volume accessed vs non-accessed in the 6M time windows starting from different data keeping periods, for the AODSIM data type.

MiniAOD

Firstly, we check the T2 level and investigate MiniAODs at T2 with respect to T1+T2. Figures 3.14 and F.7,F.9 in Appendix F, show a completely different behaviours with respect to the previous formats. If we take a look at the different time windows, we note that the total volume through 3 months to 12 months time window is increasing but this trend is followed by a decreasing of non-accessed data volume and an increasing of accessed data volume (each one roughly of 0.5PB). How the accessed data volume evolves is shown in Figure H.5 in Appendix H: we can see that in the 6 months time window the MiniAOD had a quick increase in popularity shown by the increment of all > 0 accesses bin.

Secondly, we compare the T2 level with the T1 level for MiniAODs, as in Appendix G, Figures G.7, G.8, G.9): in average the Tier-1 manages 2 PB of MiniAOD less than Tier-2; the storage is not well used as in Tier-2 for MiniAOD, only during the 12 months time window we can see a predominance of accessed data, but in more realistic 6 months we can not see a clear sign of regular/increasing use by analysts as before. This may just be explained by the fact that MiniAOD are tiny in size, and are hence hosted in T1 sites and in T2 sites, meaning that users will tend to access them in the most adequate Tier level for analysis, i.e. the T2 level, leaving the MiniAOD replicas at T1 non-accessed (or

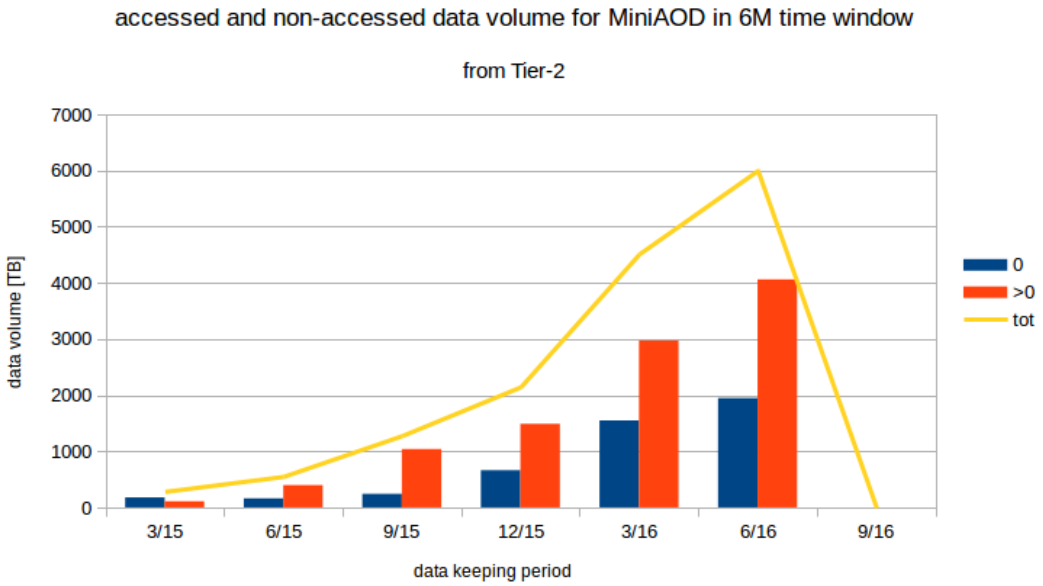

Figure 3.14: Data volume accessed vs non-accessed in the 6M time windows starting from different data keeping periods, for the MiniAOD data type.

less accessed).

All these observations go towards the conclusion that the MiniAOD is indeed a much more agile, light and adequate data tier for the CMS distributed analysis system.

At this point, aiming at condensing the observations we collected on the WLCG Tier view in this section, we display in tables below the percentage of total data volume and only accessed one, the non-accessed volume is specular, respectively for the union of Tier-1 and Tier-2, and separately for Tier-1 and Tier-2:

AOD

	T1+T2	**T1**	**T2**
total size [PB]	7.1	1.8	5.3
accessed	57%	38%	63%

AODSIM

	T1+T2	**T1**	**T2**
total size [PB]	42.5	9.7	32.8
accessed	41%	36%	42%

MiniAOD

	T1+T2	**T1**	**T2**
total size [PB]	7.2	1.2	6.0
accessed	67%	61%	68%

As shown in this table we note that effectively the Tier-2 sites balance the data accesses rate considering as average the accesses percentage related to T1+T2.

At this point, after the study of how the data access patterns evolve with time, we stop and analyse such patterns in specific time intervals, with special focus on T2 sites. As a reference, Figure 3.15 shows the number of distinct analysis users per week, i.e. individuals who submitted at least one CMS CRAB analysis job on the Grid in that week. Note that this plot goes as far into the past as September 2009. In principle we would be interested in a much shorter time window, i.e. the time window that goes from April 2014 to July 2016, so basically from the middle of LS1 to the current date. In the plot, most important holidays (only Christmas is in evidence) can be inferred from the analysis submission patterns, as well as the most relevant Summer and Winter conferences. From this plot we can see how the yearly analysis submissions pattern is marked by holidays and conferences. The hypothesis under study is that in the actual data access patterns there must hence be a cyclicity that reflect this. On the other hand, such hypothesis is hard to verify as there are many factors that may veil it, one example being the number of submitters: this number is somehow arbitrary and unpredictable, it possibly increases just before a conference, but how much it increases depends on factors very difficult to quantify, such as how relevant the conference is, what is the physics interests and "trends" of that specific moment, what is LHC data hinting the experiment communities to look into, etc). So, there is an average pattern that is somehow clear but it has a lot of structure whose nature is hard to unveil.

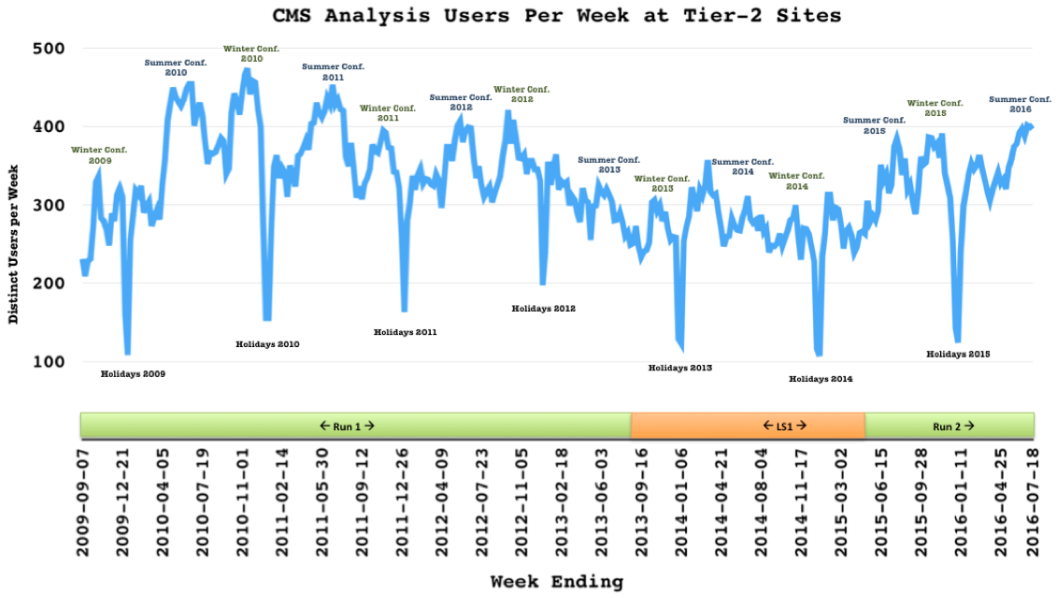

Figure 3.15: Individual CMS analysis submitters per week to the Grid from September 2009 to July 2016. [source: CMS Dashboard] [48]

Other considerations can be drawn from Figure 3.16 and Figure 3.17. The Figure 3.16 shows the number of CMS analysis jobs per week at Tier-2 sites (blue) compared with all jobs (green), from January 2015 to June 2016. On average, the number of analysis jobs was around 1.000.000 jobs per week in 2015, and in 2016 is ramping up to 1.500.000-2.000.000 jobs per week. We also know from Figure 3.15 that the minimum between December 2015 and January 2016 represents the Christmas holiday season (less than few hundreds thousands of analysis jobs submitted). Additionally, a milder correlation can be seen with Easter holidays, too. The Figure 3.17 shows the number of CMS CPU cores used for analysis per week at Tier-2 sites, excluding CERN (blue) compared with all CPU cores used (green), from January 2015 to June 2016. Comparing the two plots, the same correlation with yearly holiday seasons is visible.

Now, from both the aforementioned plot, we can take as reference the weeks between September/October 2015 and December 2015: in this period, we measured on average about 2.000.000 analysis jobs submitted and about 18.000 CPU cores used at the T2 level. In order to have more reliable data which could better describe reality, we opt as usual for a 6 months time window in the past. Therefore, in order to cover the selected period of interest (from September/October 2015 to December 2015) we must focus on a data keeping window that starts in March 2016 and go back 6 months. In this time window, firstly we analyse the volume as a function of the number of accesses to the 4 different data types at the T2 level (see Figure 3.18), and secondly we analyse the total

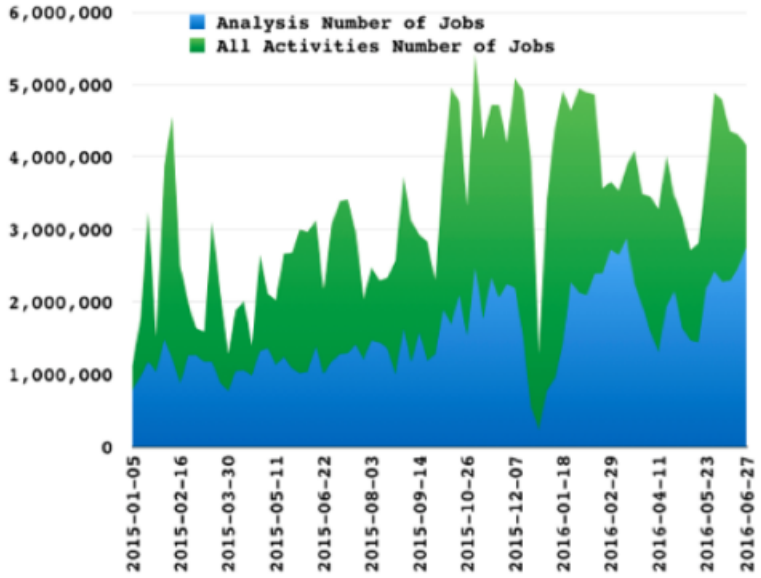

Figure 3.16: Number of CMS analysis jobs per week at Tier-2 sited (blue) compared with all jobs (green), from January 2015 to June 2016. [source: CMS Dashboard] [48]

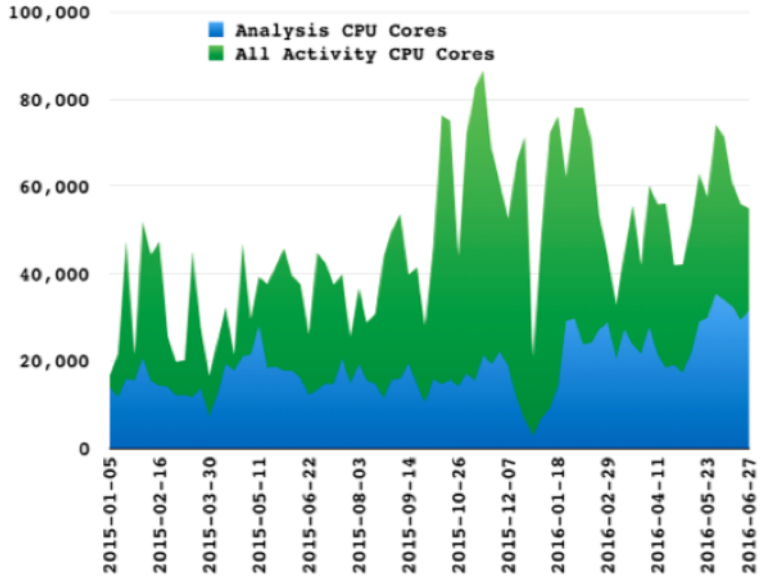

Figure 3.17: Number of CMS CPU cores used for analysis per week at Tier-2 sites, excluding CERN (blue) compared with all CPU cores used (green) during the same time interval. [source: CMS Dashboard] [48]

used (> 0 accesses) and not used (0 accesses) data volumes for the 4 different data types at the T2 level (see Figure 3.19).

In Figure 3.18, the scrutiny plot for the selected period is shown, highlighting the various data type contributions in different colours. For AOD one can see that the 0 old bin is comparable to the > 14 bin, and looking ad the other bins > 0 one notes that this format is always used; the 0 bin counts almost 0.5PB in this period it means that only 0.5PB of new data have 0 accesses but just look at the 1 access bin that counts 0.8PB. The AODSIM dominates the first three bins and that recalls some early considerations: it is the most numerous data type, its production is really quick in fact neither in 6 months we can lower the 0 bin that still counts 5PB, the oddness of its 1 access bin. Again, the MiniAOD format displayes its quality, just look at the 0 old bin the lowest of all format; the 0 bin is quite the same of 1 access bin (respectively 1312TB and 1306TB) those aren't so far from 759TB of > 14 accesses bin. Finally, there is the RECO data type, that considering its total volume, shows the highest 0 old bin and its presence in other bins ¿0 is not relevant and that seems to support our thesis about this format disuse. Now, in Figure 3.19 we can see that the AODSIM format is the most used and also the most envolved but it isn't as convenient as MiniAOD; this latter involves less space than AOD and AODSIM and performs the same analysis of those two, so we can say that the best resource utilization is realised by MiniAOD and the worst one by RECO - more than half of total volume is unused and it is composed in great part by really old data (0 old).

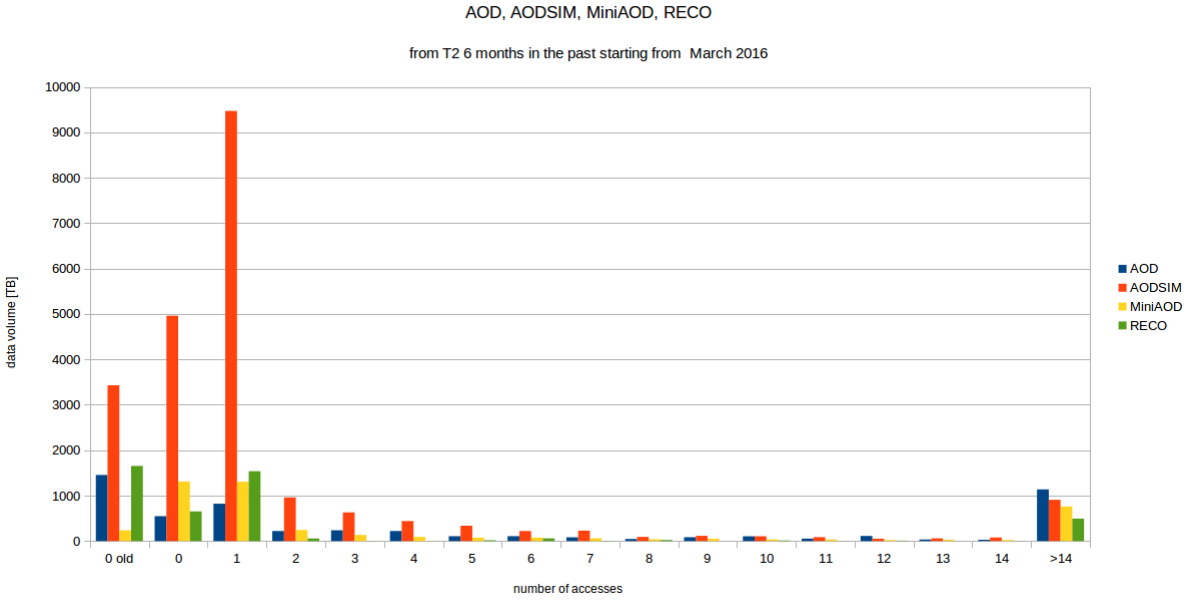

Figure 3.18: Data volume accessed in terms of number of accesses, with breakdown into the 4 different data types at the T2 level for the last quarter of 2015.

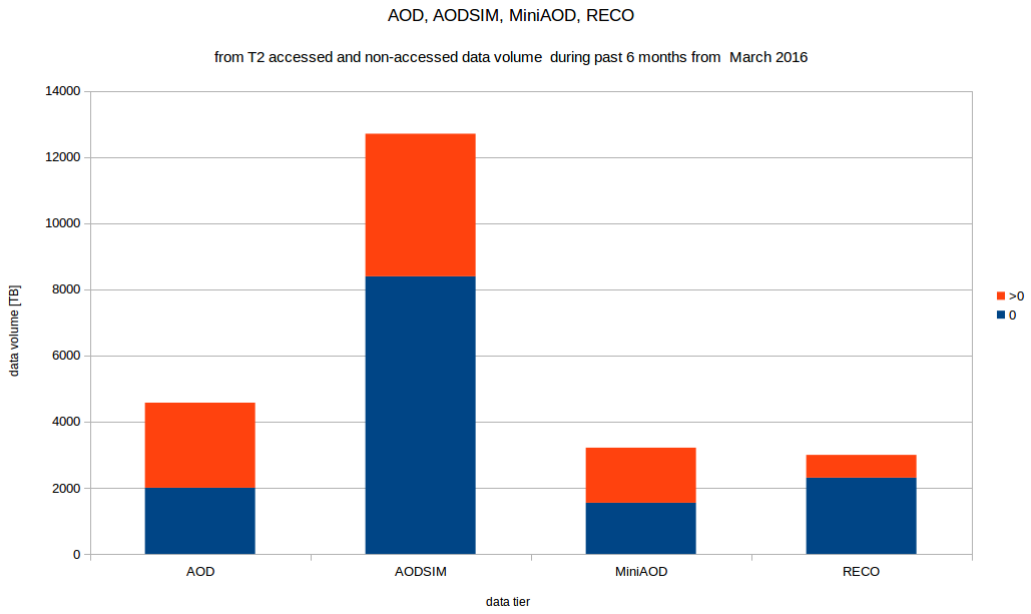

Figure 3.19: Data volume accessed vs non-accessed, with breakdown into the 4 different data types at the T2 level for the last quarter of 2015.

Chapter 4

Application of Machine Learning techniques

In several areas of the computing model, the CMS experiment can take advantage of Machine Learning approaches. In this chapter we will discuss the work done to attack the problem discussed thoroughly in Chapter 4 with ML techniques.

4.1 Introduction on Machine Learning

Machine Learning (ML in the following) is one of todays most rising technical fields, and recently a revived buzzword in the advanced computing techniques. Lying between computer science and statistics, ML is at the core of a discipline now called "data science and of entire sectors of study like Artificial Intelligence (more below). In a nutshell, ML is one of the answer about how to build computers that improve through experience. The adoption of ML methods covers a wide range of application, ranging from science and technology to commerce, even to decision-making processes such as health care, education and marketing. It is used as a tool to perform tasks behind the behaviour of every web search engine like Google, for example. An interesting example is Facebooks News Feed, which uses ML even in real-time to personalize each members feed: if a member frequently stops scrolling in order to read a particular friends post, the News Feed will start to show more of that friends activity. To do so, the software is using statistical analysis and predictive analytics to identify patterns in the user data and use those to populate the New Feed. As said earlier, ML is also closely related to Artificial intelligence, Computational Intelligence and Data Mining fields. Artificial intelligence is a branch of computer science, focussed on the study of computational system that do things that men can do or to do things evaluated as intelligent'. The connection is the learning process, a feature of an intelligent system, and ML is concerned with the study of systems capable of learning. Computational intelligence is instead concerned with

systems complex behaviours: this field produces system for subfields such as artificial immune systems and artificial neural networks. The learning process, in this case, starts from interactions with their environment. Finally, Data Mining includes systems that discover relationships among large data sets. It is also known as knowledge discovery in databases. Data mining is endowed in computer science, and in this case the contribution from ML is a set of tools to learn relationships in data that provide the basic discovery.

As a general concept, therefore, the emphasis of ML is on automatic methods, and its paradigm can be stated as easily as "learning by examples". A widely accepted definition of this is:

A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, ad measured by P, improves with experience E.

For instance, a computer program that learns to play checkers might improve its performance as measured by its ability to win at the class of tasks involving playing checkers games, through experience obtained by playing games against itself. In this complicated process, an essential step is to define well the learning problem; again, in the previous example:

- Task T: playing checkers
- Performance measure P: percent of games won against opponents
- Training experience E: playing practice games against itself

4.1.1 Learning techniques and problems

The main characteristics of modern ML are:

- The primary goal is to gain highly accurate predictions on test data;
- The methods should be general-purpose, fully automatic and off the shelf (with a prior human knowledge);
- Emphasis should be on methods that can handle large datasets.

The main learning techniques used in all applications are Supervised ML and Unsupervised ML, briefly described in the following.

4.1.2 Supervised and Unsupervised Machine Learning

Supervised ML

Supervised ML is based on adaptive algorithms which - through observation - identify patterns in data with the aim to build a model that makes predictions within evidence. In this case, the learning algorithm is fed with an initial known data (input) and exact correspondent results (output). Then, it is possible to build a model that will be tested to generate reasonable predictions for the new output data. Examples of supervised learning are:

- *Classification*:the goal is to assign a class (or label) from a finite set of classes to an observation. Any classification model consists in two datasets: training set and testing set. This can be thought as a discriminant problem: starting from known labelled data, making a model then looking for a pattern in dataset ("train"), finally make prediction about which label will be assigned to new data (test). An algorithm used for classification problem is the "decision tree: here data population is splitted by a feature which most evenly divides the set. Final output ("label") is the result of many subsequent splitting, as above.

- *Regression*: it manages data labelled with real value rather than a label, for instance: cost of houses, total sales etc. (in classification problems, the data type is heterogeneous), Here, one establishes a relationship between independent and dependent variables by fitting the best line. In Linear Regression, for example, the best fit line is given by Y = a * X + b, the coefficients a and b being computed by minimizing the sum of squared difference of distance between data points and regression line. Then, the problem is: given a new X value which will be the most probable value of Y?

Unsupervised Machine Learning

Unlike Supervised ML, in Unsupervised ML there are no initial output variables corresponding to input data. The goal in this case is to model the underlying structure in the data. The main Unsupervised ML problem is:

Clustering: A cluster is a collection of data items which are similar between them and dissimilar to data items in other clusters. An important feature is distance. All cluster analysis hinges on the idea of two things being close in a descriptive sense within the data space, this latter is called dimensionality. Dimensionality can directly impact on distance measurements; then in clustering analysis we have to threat these dimensions ad features (item purchases, age, income etc.), distance on these feature can be many things (euclidean for instance). Another important step is to reduce dimensionality to simplify the problem: some features are simply proxies to other features. Approaching to clustering useful algorithm types are Centroid-based, K-means, for example.

4.2 Applications of Machine Learning in CMS

ML techniques are applied in CMS in several areas, the description of which goes beyond the scope of this work. In the development of this thesis, though, it was conceived since the very beginning to exploit some ML approaches to set-up a machinery to get prepared to predict the future popularity of the most interesting CMS data type for analysis. The results in Chapter 4 showed that there is a data type that is extremely promising for CMS in terms of efficient storage utilization, and this is the MiniAOD format. In the following of this chapter we hence describe the work done to set-up a machinery to perform MiniAOD popularity predictions.

4.2.1 The CMS DCAF machinery

A previous work done in the CMS Bologna group and documented in publications and in a thesis [49][50] was exploited to progress in this study. In connection with the central CMS teams focussing on R& D on Big Data and Analytics techniques the CMS-Bologna team adapted and utilized a so-called DCAF pilot machinery to apply ML algorithms to old popularity data in order to predict future popularity data. A full description of DCAF goes beyond the scope of this thesis, but a full description can be found here [49]. Only major and most relevant information is summarized in the following.

The DCAFPilot (Data and Computing Analysis Framework Pilot) [51] is a pilot project that offers a framework to apply ML algorithms. The architecture is such that the data is collected from CMS data services by a DCAF core component that uses MongoDB as a technology for its internal cache. Such informations are collected from several CMS structured data-services, namely DBS, PhEDEx, PopDB, SiteDB, Dashboard [52]. A dataframe generator toolkit has been designed and developed to collect and transform data from the aforementioned CMS data services, and to extract all necessary bits for a subset of popular and un-popular datasets. The dataframe is then fed to standard ML algorithms (both python and R code is used for this) for data analysis. A quantitative estimate of the popularity can be given for specific types of datasets, which may be eventually fed back to the CMS computing infrastructure as a useful input to daily operations and strategical choices.

4.3 Application of ML to the CMS data access study

In this thesis we focus on the set-up of the DCAF machinery to attack the problem of the prediction of MiniAOD popularity in CMS as a classification-like Supervised ML problem. We create a model and train it on a training set of data from the past, we compare predictions given by the model with the valid set of data - still from the past

but independent for the first set - and we set-up all necessary step to further tune the model to perform actual predictions. Results about the quality of the model we can build may be shown in different manners, we opt for a very simple and traditional one, e.g. a standard confusion matrix plus an accuracy scorer. The confusion matrix is a table with two rows and two columns that reports the number of false positives (FP), false negatives (FN), true positives (TP), and true negatives (TN) predicted by the model. The entries in a confusion matrix are counts, i.e. integers (see Fig.4.1).

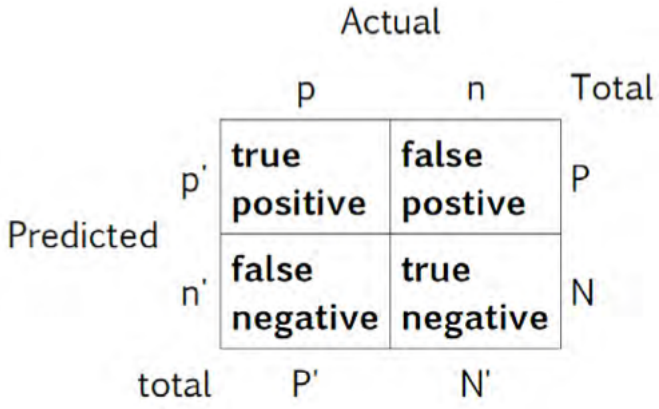

Figure 4.1: Confusion matrix (see text for explanations).

The total of the four entries TP+TN+FP+FN=n is the number of train examples being used. We will use a Random Forrest classifier as ML classifier (a classifier is an algorithm that implements the ML classification). In order to evaluate its outcome in the following considerations we may opt for several estimators (Accuracy, Precision, Recall, F1) but we focus on just Accuracy in the following, for simplicity. So, results and considerations will be made talking of values of TP, TN, FP, FN and accuracy of the model:

- TP integers are related to how many MiniAOD datasets are predicted by the model to be popolar and turned out indeed to be popular ;
- TN integers are related to how many MiniAOD datasets are predicted by the model to be not popolar and turned out indeed to be not popular ;
- FP integers are related to how many MiniAOD datasets are predicted by the model to be popolar but turned out to be not popular instead;
- FN integers are related to how many MiniAOD datasets are predicted by the model to be not popolar but turned out to be popular instead;
- Accuracy is a percentage value that indicate how good the model we are building is.

Given all these important premises, in the follwing we describe the tests done and the results achieved, in subsequent stages.

4.3.1 Blind application of a general model

First of all, we tried out the standard ML approach used in [49], in which a model was tuned on (mostly) 2014 data to predict analysis data tiers popularity. Note that in this case all data tiers used in analyses were considered altogether with no distinction, i.e. AOD, AODSIM, MiniAOD, RECO. Note also that in the training period the MiniAOD just started to appear in the CMS managed data. Additionally, the definition of popularity used is based on naccess > 10 (number of accesses to the datasets greater than 10) and totcpu > 10 (number of hours spent on CMS CPUs to access the data greater than 10 hours), which was found to be the optimal tune for the approach and needs of [48]. The only modification applied was to train the model on MiniAOD only and see the model quality outcome. Admittedly, such model is completely inadequate to be used for the present problem, but a first, blind run of that model on MiniAOD only allowed to refresh the tool and adapt it to run on MiniAOD. The outcome was, as expected, a complete failure, the model hardly reaching an accuracy of 40%. We then moved to the second step.

4.3.2 Ad-hoc training of a better model

First of all, the data on which the model training is done need to include as many experience on MiniAOD as possible. Optimal would have been to exploit all data from early 2014 onwards, until today. Unfortunately, it takes plenty of work to prepare the data in input for such process and at the time of this thesis the work is still on-going by CMS collaborators. The most recent data we could exploit for this study go from early 2014 up to May 2015 (excluded). It is not optimal but also not extremely bad, and it deserves a try. MiniAOD existed already for some months, despite their access pattern may not have been as high as in 2016. Also, one need to consider that not the 100% of data is used for training: we need - following a common ML practice - to divide the sample into a training set and a validation set, the first being the one on which one builds the model and the second being the one, independent from the first on which the model is checked for its predictive capabilities, and TP, TN, FP, FN, accuracy estimators are computed. In general one should divide the entire data sample into a 70%-30% for training and validation respectively: we used January 2014 to February 2015 for training, and March 2015 to April 2015 for validation. On this set-up, we ran the ML as from the DCAF framework again on a multiple set of possible definitions of popularity. We exploited naccess$> N$ with N ranging from 1 to 10, totcpu$> M$ with M ranging from 1 to 10, and we also added tests on an additional variable that was discarded in the popularity definition as of [49] but we thought would have been interesting in this case, i.e. nusers

(number of analysis users accessing a given MiniAOD dataset), with cuts as of nusers $> P$, with P ranging from 1 to 5.

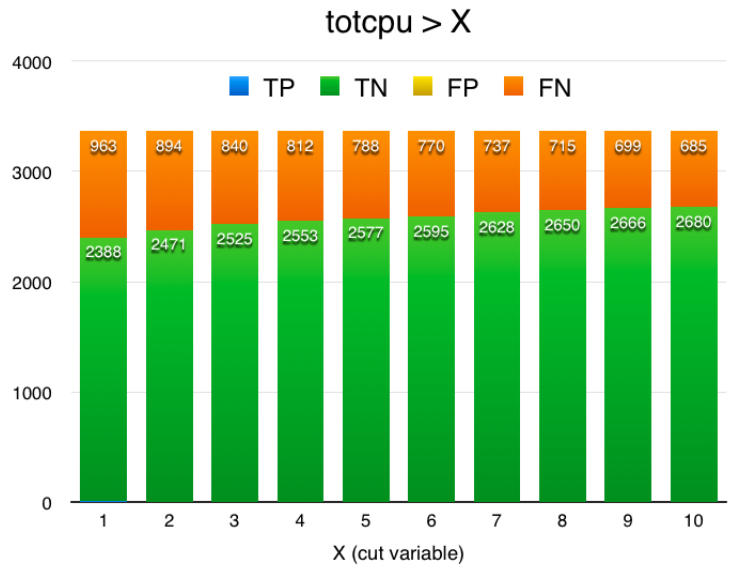

Figure 4.2: Results for the accuracy estimators applying the totcpu as cut variable.

The results of the application of just the single cut on totcpu$> X$ are show in Figure 4.2. It can be seen that there is no obvious choice of the cut that allows to achieve a vast majority of TP and TN: we observe, among the 3365 datasets considered, a vast majority correctly predicted as non-accessed (TN, i.e. the training said they would not have been accessed in the validation period, and indeed they werent) with a minority wrongly predicted as not-accessed (FN, i.e. they were predicted not to be accessed but they were accessed instead). Due to unavailability of more recent data, the training set covered a period in which MiniAOD either did not exist yet or existed and were randomly and lightly accessed still: any model trained on this sample will predict 0 or close-to-0 accesses to the sample in the following validation period (indeed, only TN and FN integers have been reported by the model), and indeed in the validation period the MiniAOD usage started to grow more and more so all outcome is negative (i.e. no access) but actually some were instead accessed (i.e. the FN fraction). Actually, despite not visible in the plot, there is one cut that allowed also to see some TP: with totcpu¿1, the model reported 14 TP and 2388 TN: this may hint that an integer value for totcpu is not adequate in this MiniAOD case, and the code in the model should be changed to allow real values, to include also analysis jobs that may use < 1 hours to access MiniAOD on CMS CPUs. The accuracy of the model with cuts on totcpu goes from 71% (totcpu¿1) to 80% (totcpu¿10).

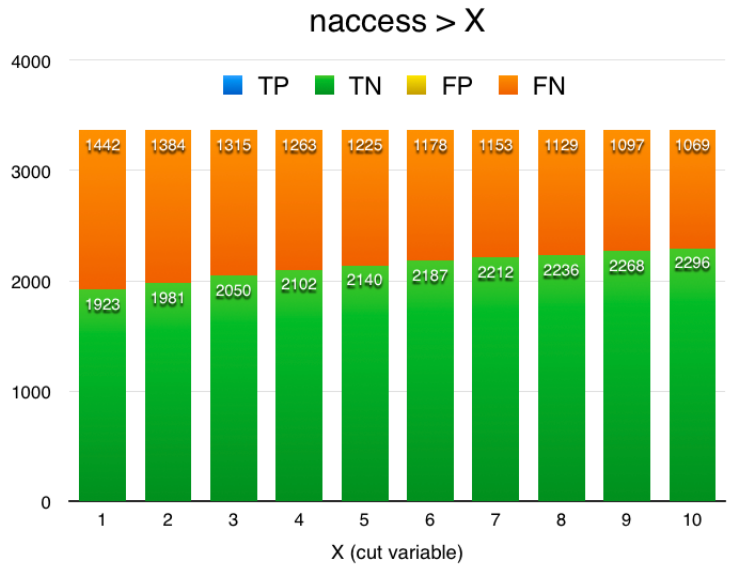

Figure 4.3: Results for the accuracy estimators applying the naccess as cut variable.

The same test has been performed for a single cut on naccess$>X$ and the outcome is shown in Figure 4.3. Same conclusions can be drawn. Only difference is that the accuracy of the model with cuts on naccess goes from 57% (naccess>1) to 68% (naccess>10), i.e. slighly lower that the totcpu cut.

As a last investigation, we performed the same tests also on a single cut on nusers$>X$ and the outcome is shown in Figure 4.4. In general, in terms of true vs negative predictions, same conclusions can be drawn here as well. But it is remarkably evident that the accuracy of the model with cuts on nusers goes from 74% (nusers>1) to 94% (nusers>5), i.e. much higher than other single cuts, and with a different slope. The model accuracies achieved in these 3 single-cut tests are shown in Figure 4.5. It can be seen that a single cut may not be the nest approach and a set of combined cuts should be explored to find a combination that defines the popularity in this specific case in a way that maximizes the model accuracy. This goes beyond the scope of this thesis and would anyway not tactical to perform all these tests on combined cuts before a more complete training period is availble to build a better model than the existing one.

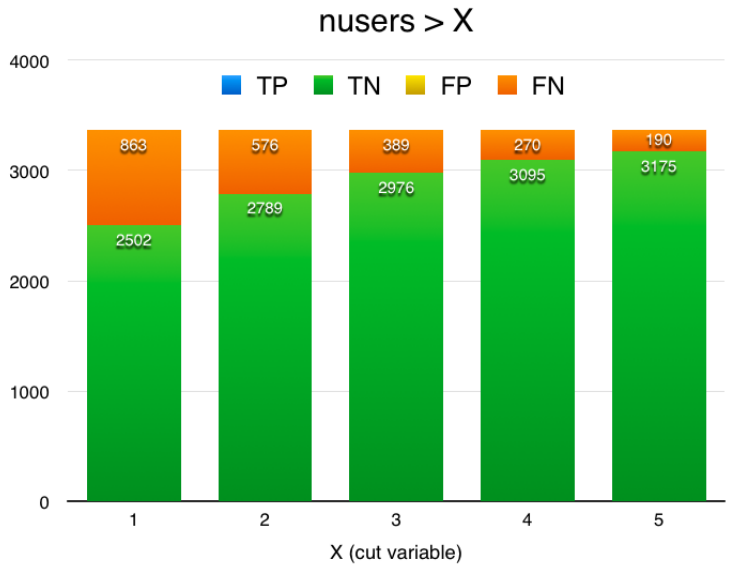

Figure 4.4: Results for the accuracy estimators applying the nusers as cut variable.

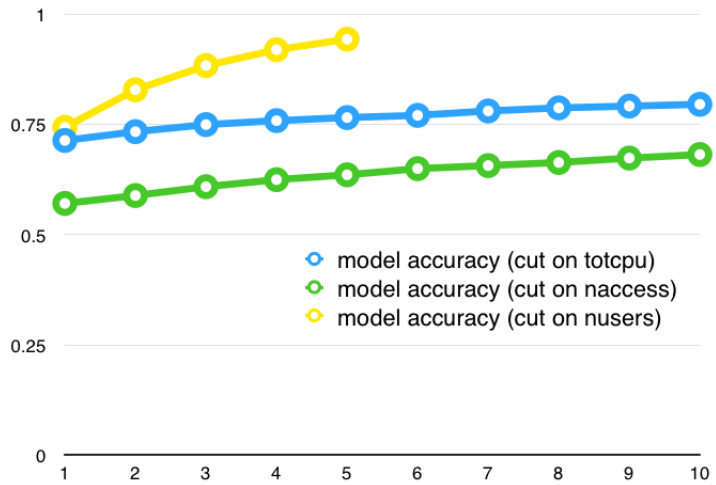

Figure 4.5: Comparison between model accuracy obtained applying the three different cuts: totpcu, naccess, nusers.

4.3.3 Towards the best possible model

The good information coming from this test is that the model based on DCAF just works and is able to produce meaningful results: it is light, easy to run and tests can be extended very easily to this problem, now that this thesis has clearly highlighted the areas of focus that are the most promising (i.e. MiniAOD studies). The current status is that we need to wait and gain access to more recent data on CMS accesses to MiniAOD in 2015-2016 and in the adequate dataframe format to be ingested by DCAF before drawing further conclusion: this recent data sample will allow to extend the training to a period in which MiniAOD started to be more heavily used, and hence the tests on training/validation sets in this case will allow to build a better model, on which a proper work on cut tuning can be scheduled and performed.

Chapter 5

Conclusions

This thesis focussed on a study of the Grid data access patterns in distributed analysis in the CMS experiment at the LHC accelerator.

This study started from the detailed evaluation of the data that are collected from CMS databases and go into the standard plots to the WLCG Computing Resources Scrutiny Group (CRSG). A deeper analysis on those informations castes light on the relative contributions that specific CMS data types give to the overall observed behaviour. In brief, the large amount of not so frequently accessed AOD and especially AODSIM data types is the major responsible for the inefficiencies in the exploitation of the storage capacities available at WLCG Tiers for distributed analysis. Regarding other formats, the RECO data type is known to be in the process of becoming a transient format, and the new MiniAOD data type emerged as the most promising one in terms of same physics content, agility thanks to its limited size and great potential - thanks to the low size - to be the basis of more healthy and flexible CMS data placement in the future. Additionally, the analysis confirmed that the CMS data access patterns is heavily dependent on the data type and not from the WLCG Tier level, thus confirming with evidence a process that is on-going since Lng Shutdown 1 in CMS, i.e. that the boundaries among the Tier-1 and Tier-2 levels are becoming more and more blurry in the CMS Computing Model.

Focussing on the MiniAOD format only, in this thesis the set-up of the DCAF machinery - a framework available in CMS to apply Machine Learning techniques to various problems, to which the CMS-Bologna group contributed even prior to this work - was used to explore the feasibility of a prediction of MiniAOD popularity in CMS as a classification-like Supervised Machine Learning problem. The results showed that this approach is possible and produce meaningful and consistent results. More statistics is needed to train the model appropriately, and to ultimately produce MiniAOD access predictions that could be used in production: this will be the content of further studies.

Appendix A

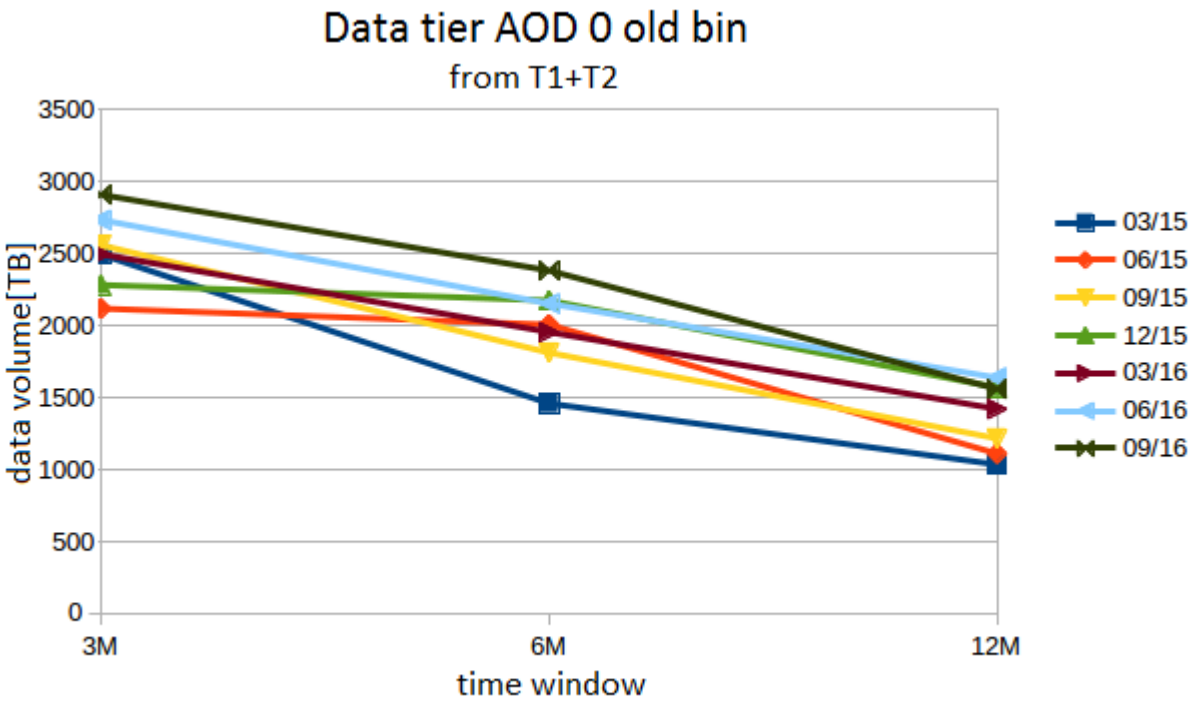

Figure A.1:

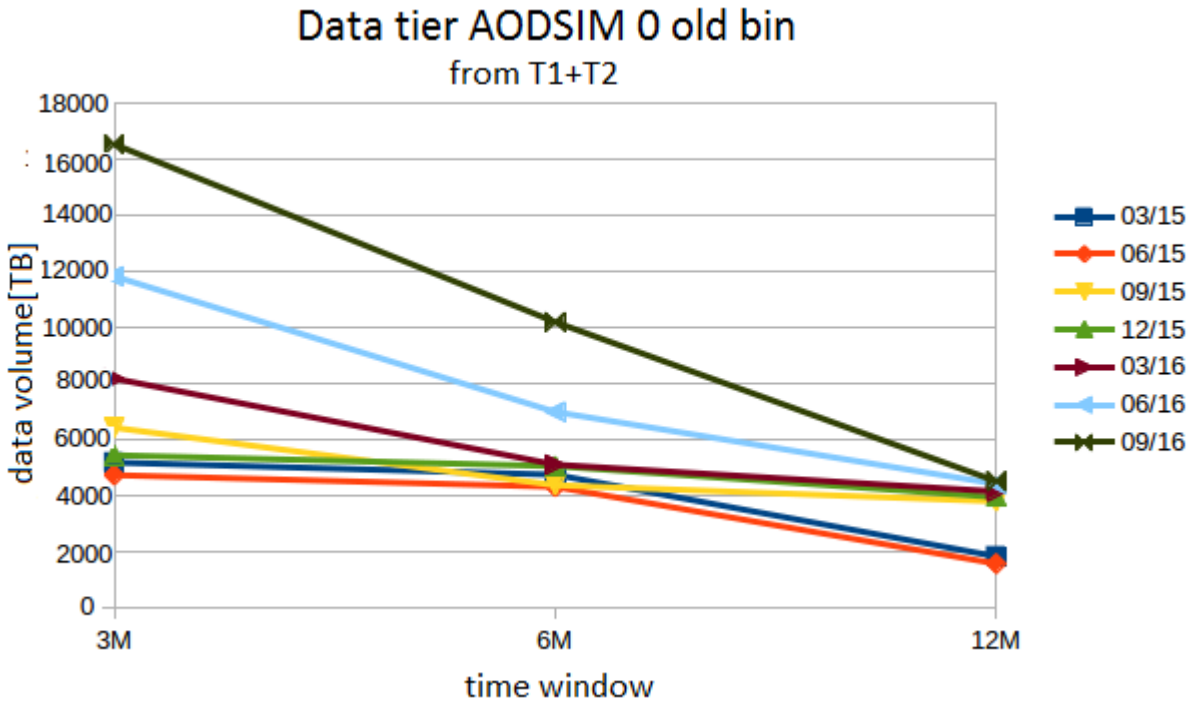
Data tier AODSIM 0 old bin
from T1+T2
data volume[TB]
18000
16000
14000
12000
10000
8000
6000
4000
2000
0
3M
6M
12M
time window
03/15
06/15
09/15
12/15
03/16
06/16
09/16

Figure A.2:

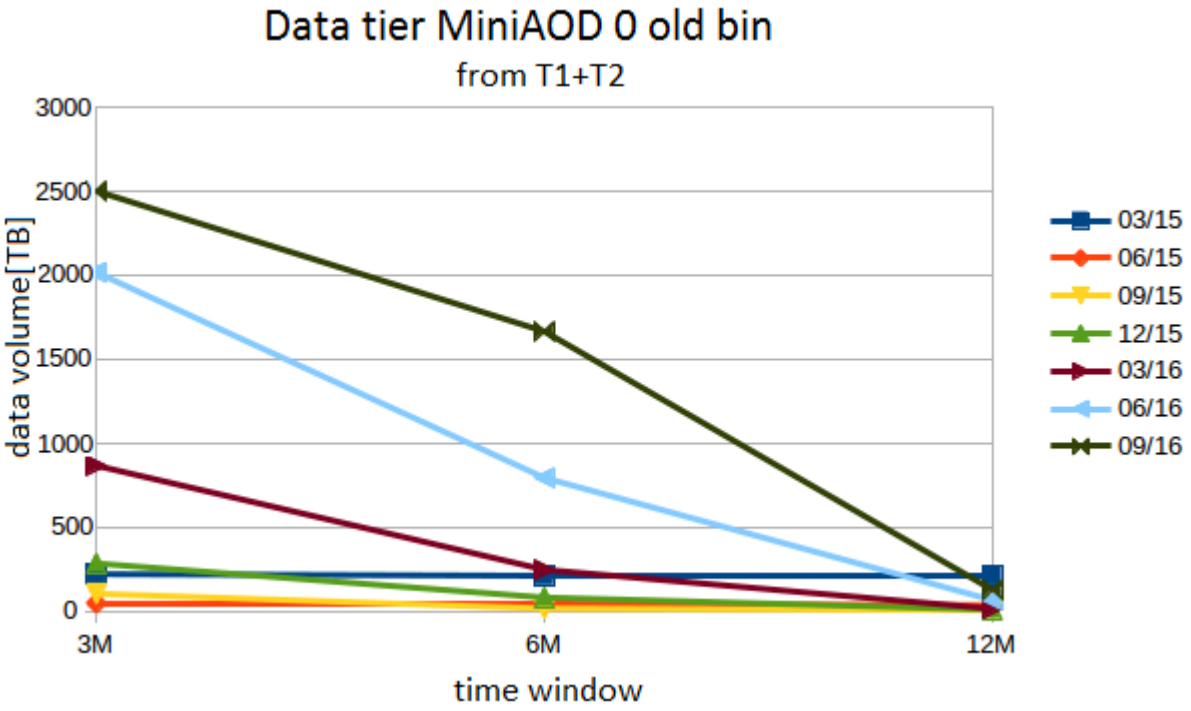
Data tier MiniAOD 0 old bin
from T1+T2
data volume[TB]
3000
2500
2000
1500
1000
500
0
3M
6M
12M
time window
03/15
06/15
09/15
12/15
03/16
06/16
09/16

Figure A.3:

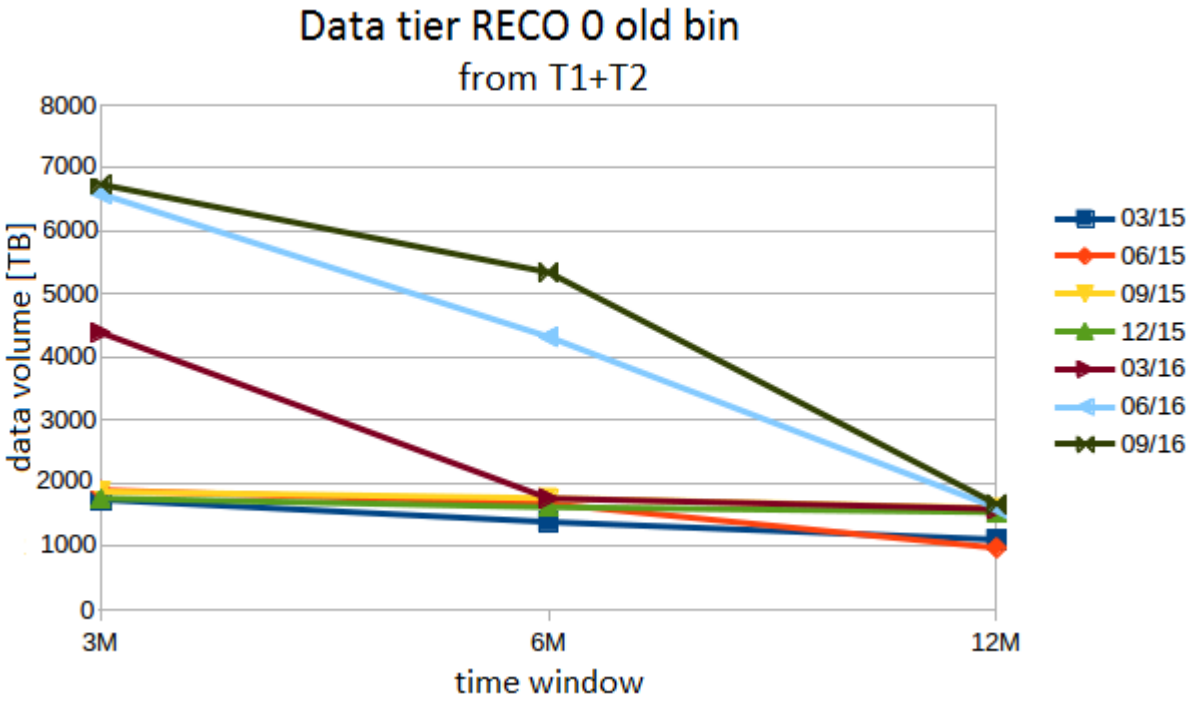
Data tier RECO 0 old bin
from T1+T2
data volume [TB]
8000
7000
6000
5000
4000
3000
2000
1000
0
3M
6M
12M
time window
03/15
06/15
09/15
12/15
03/16
06/16
09/16

Figure A.4:

Appendix B

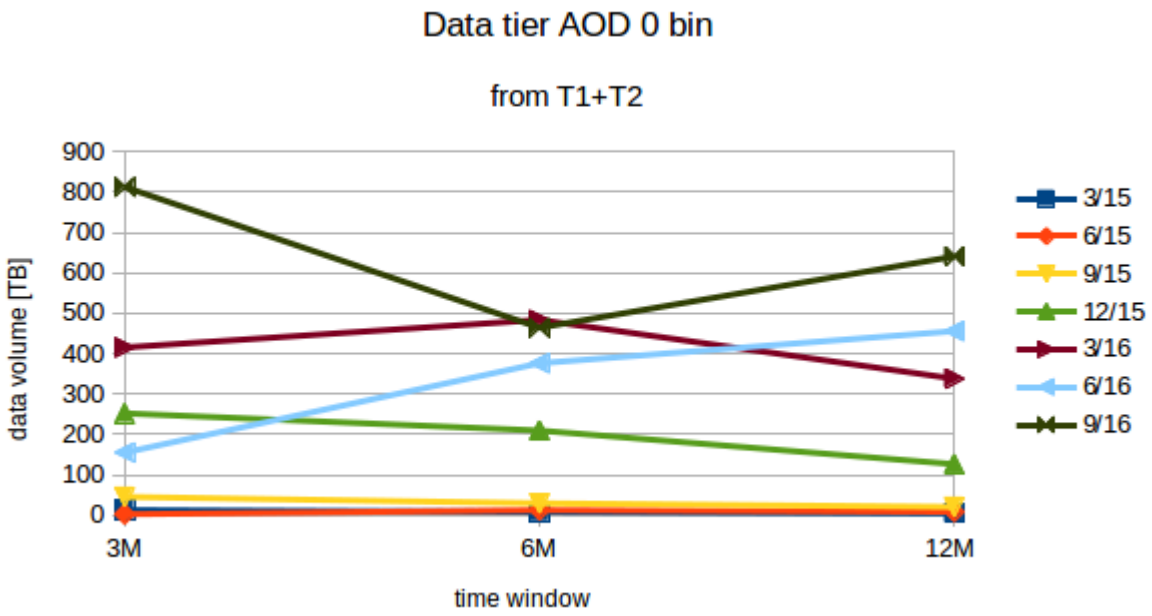

Figure B.1:

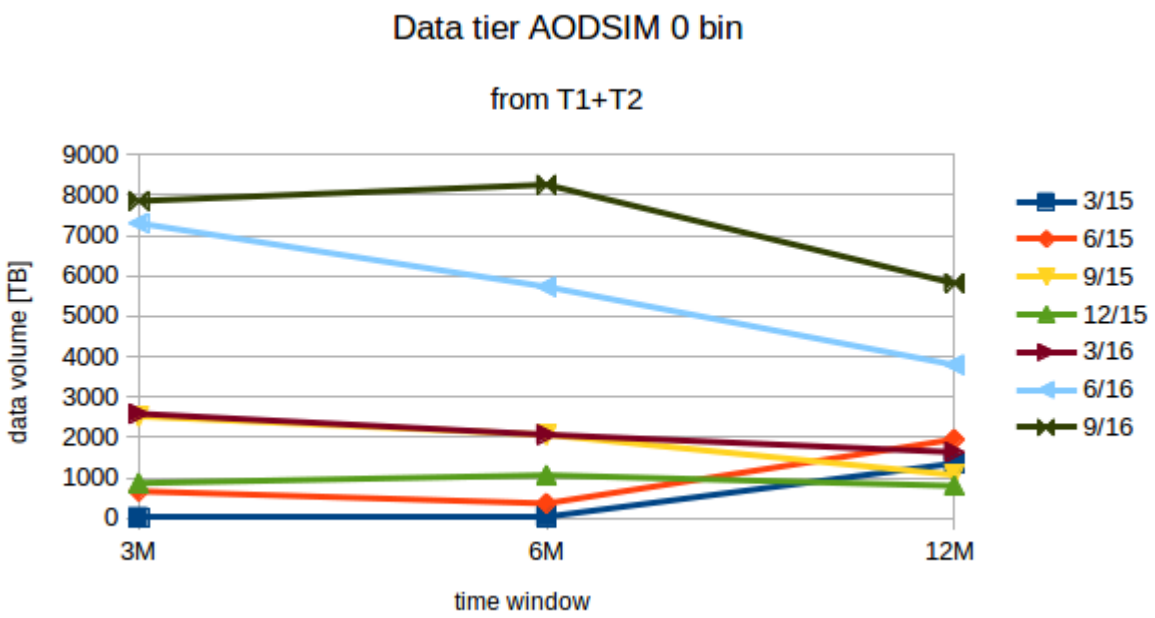

Figure B.2:

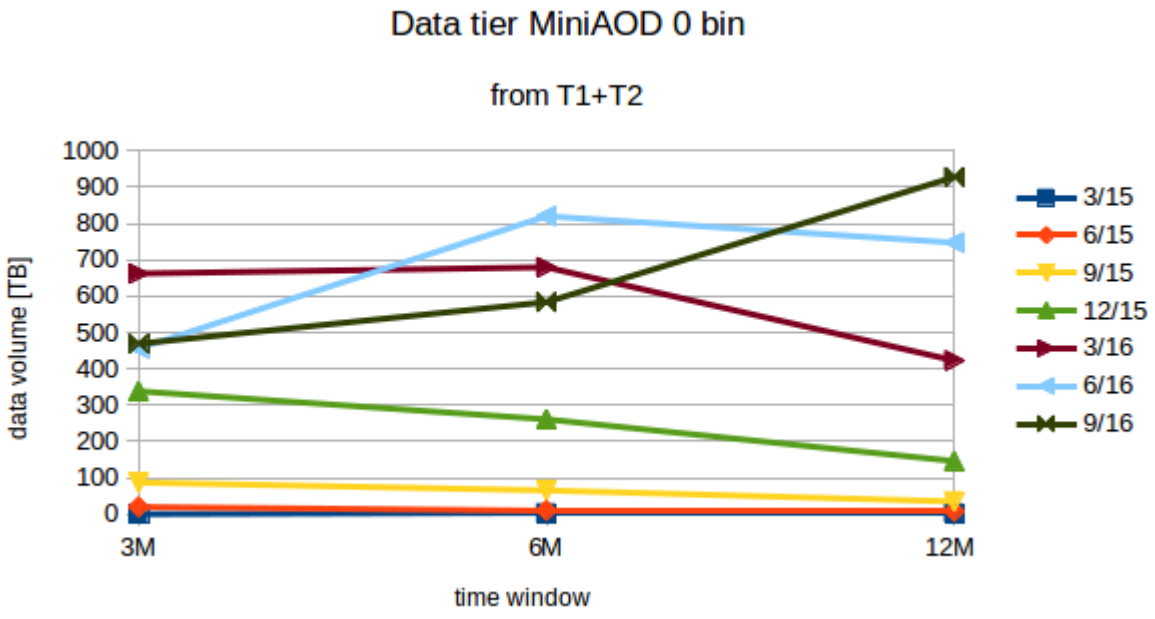

Figure B.3:

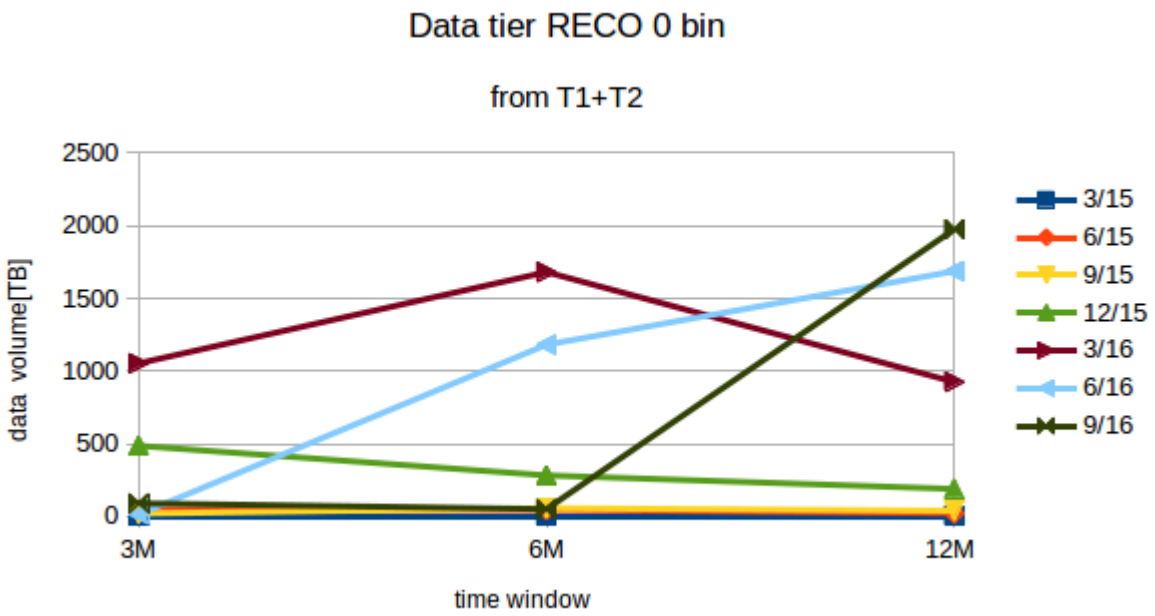
Data tier RECO 0 bin
from T1+T2
2500
2000
1500
1000
500
0
data volume[TB]
3M
6M
12M
time window
3/15
6/15
9/15
12/15
3/16
6/16
9/16

Figure B.4:

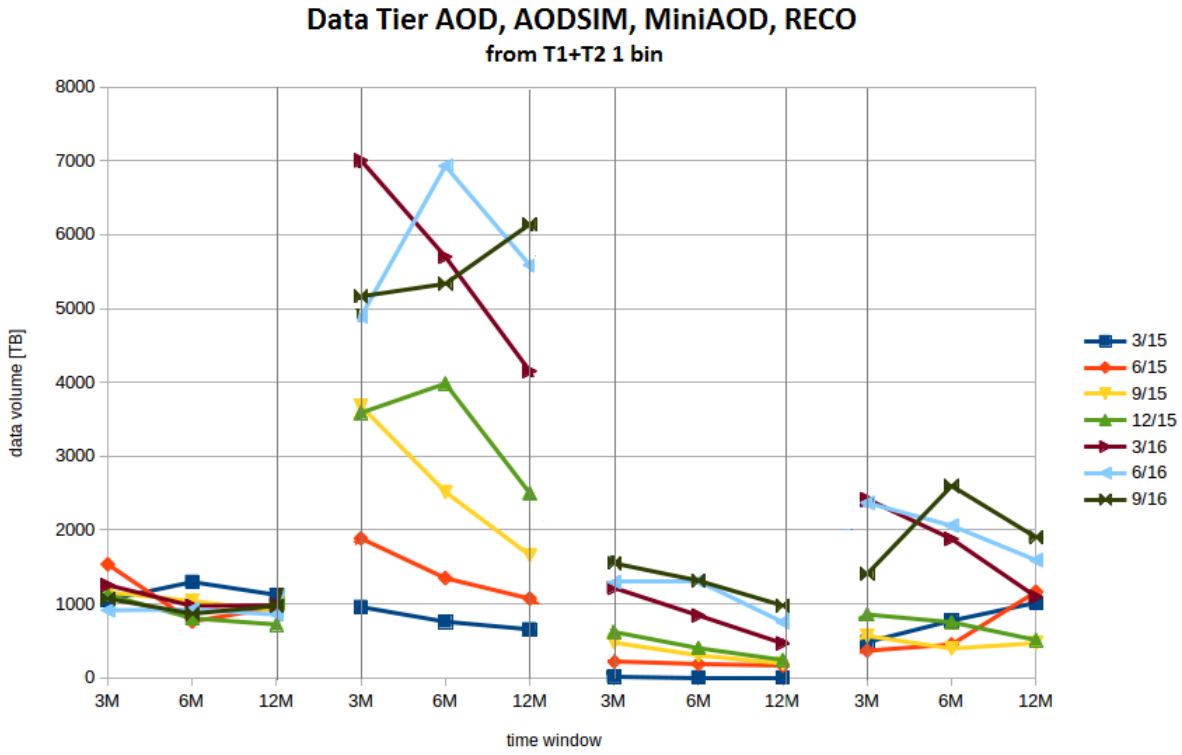
Data Tier AOD, AODSIM, MiniAOD, RECO
from T1+T2 1 bin
data volume [TB]
8000
7000
6000
5000
4000
3000
2000
1000
0
3M
6M
12M
time window
3/15
6/15
9/15
12/15
3/16
6/16
9/16

Figure B.5:

Appendix C

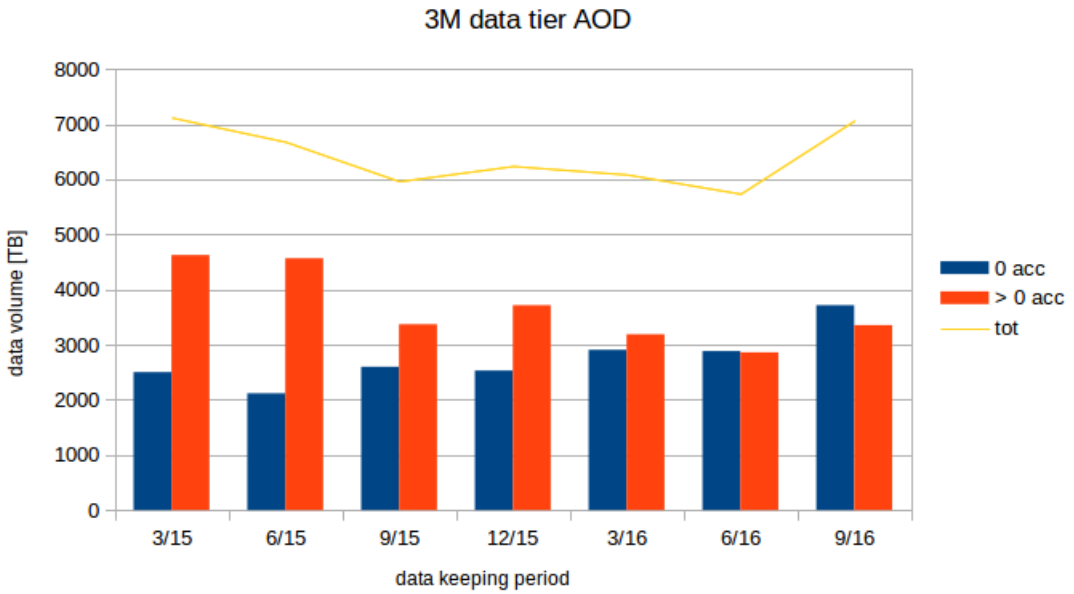

Figure C.1:

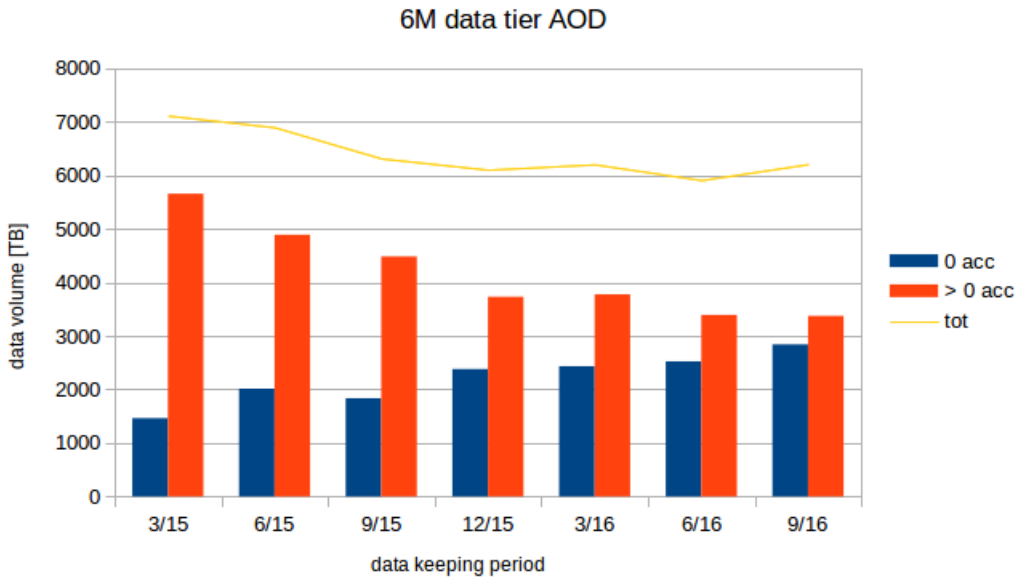

Figure C.2:

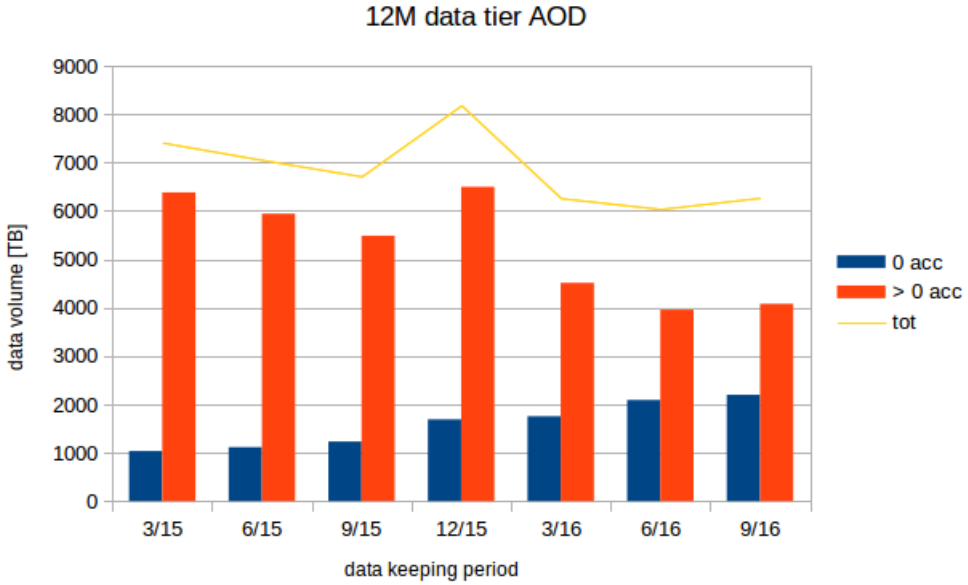

Figure C.3:

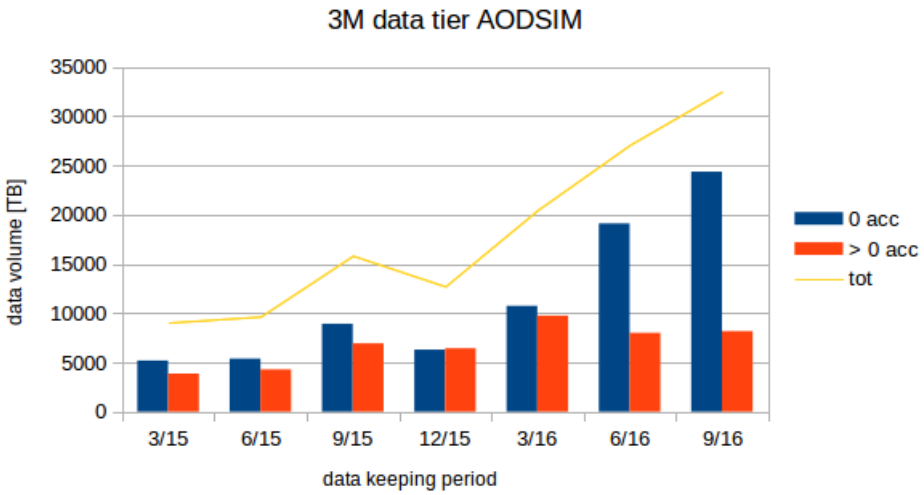

Figure C.4:

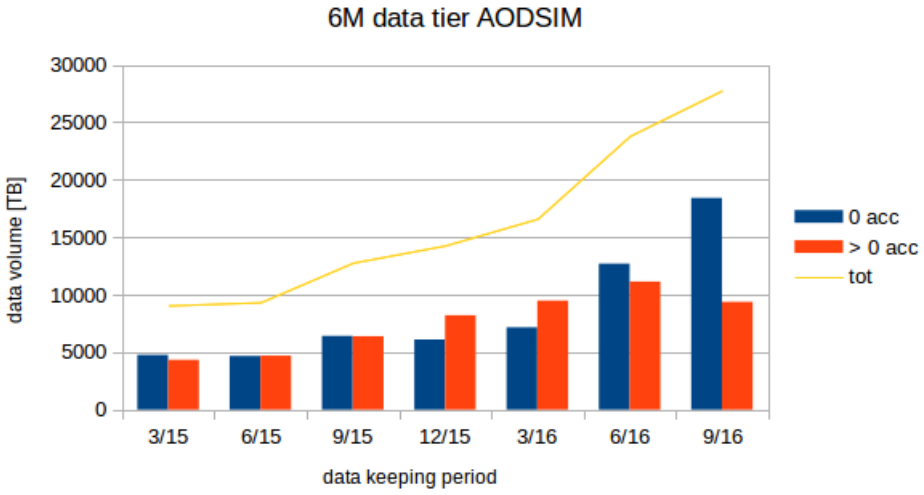

Figure C.5:

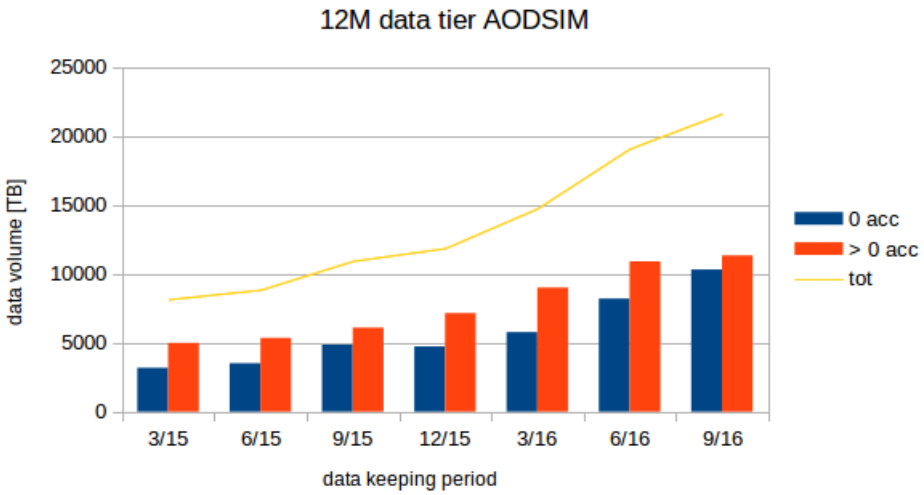

Figure C.6:

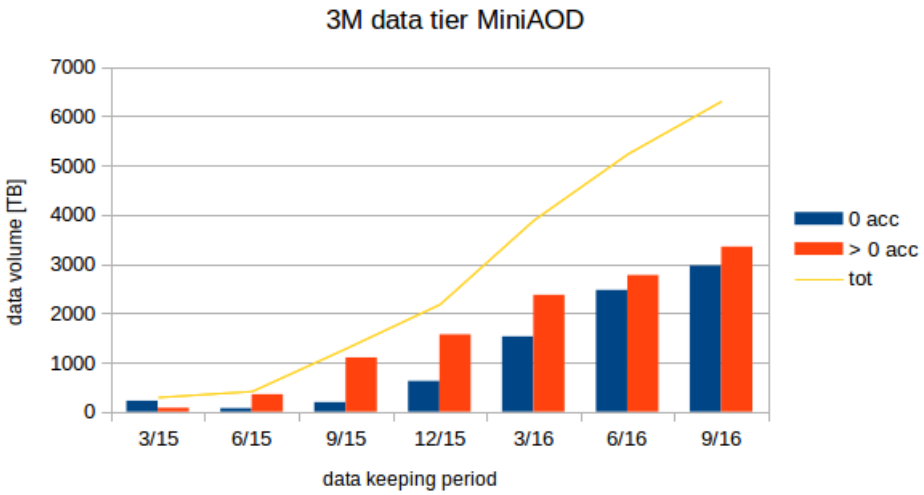

Figure C.7:

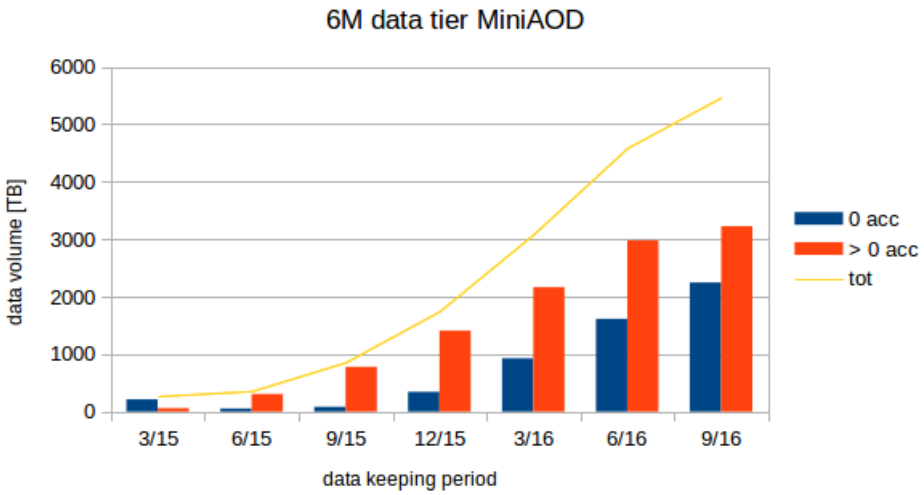

Figure C.8:

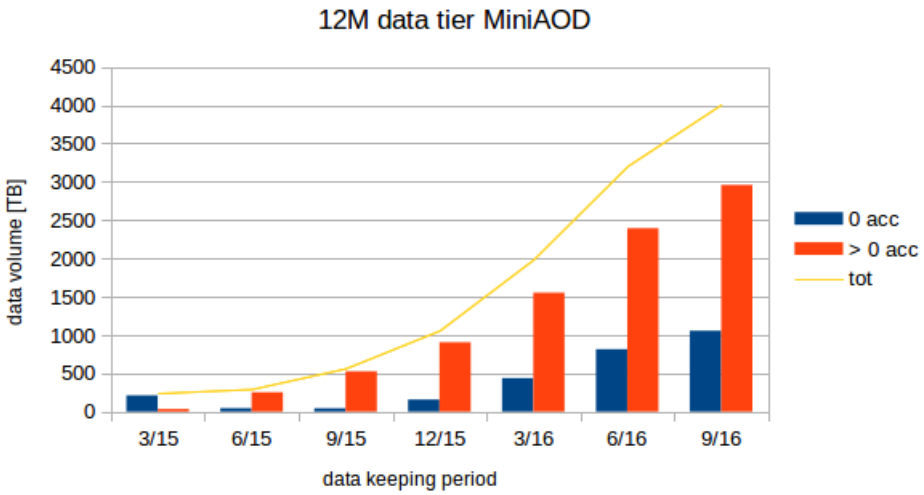

Figure C.9:

Appendix D

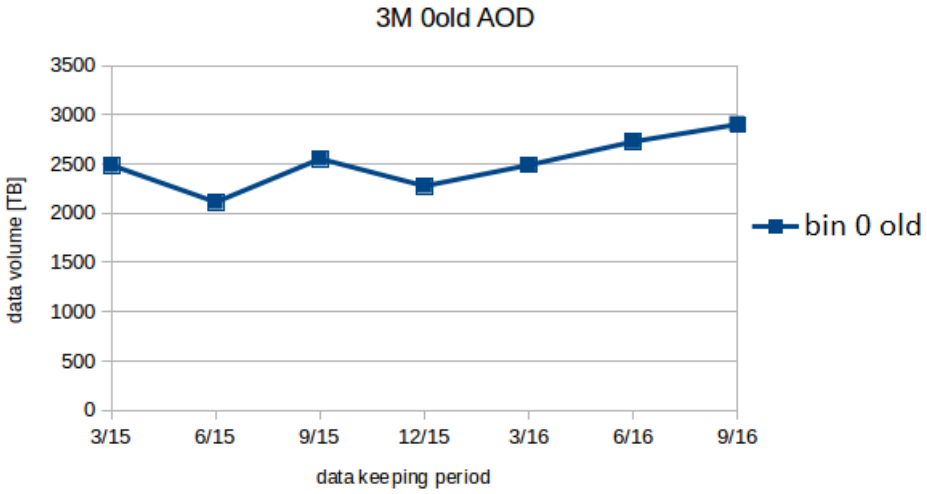

Figure D.1:

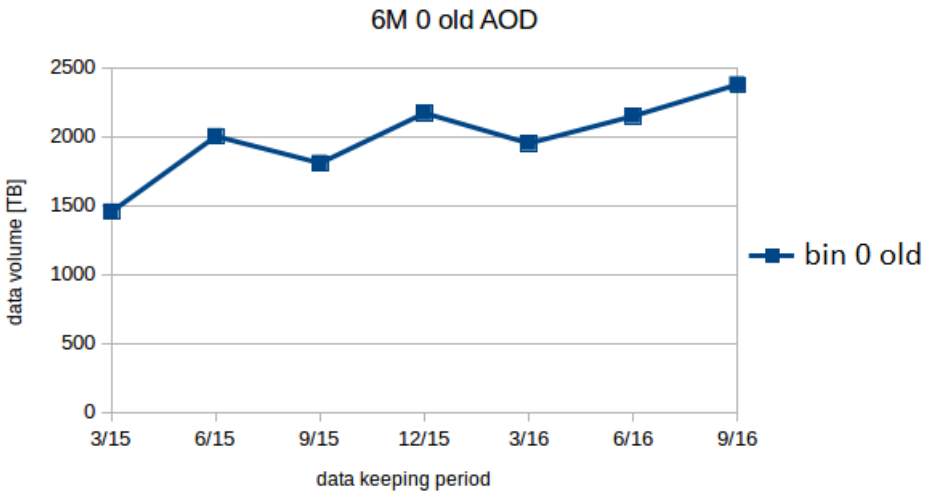

Figure D.2:

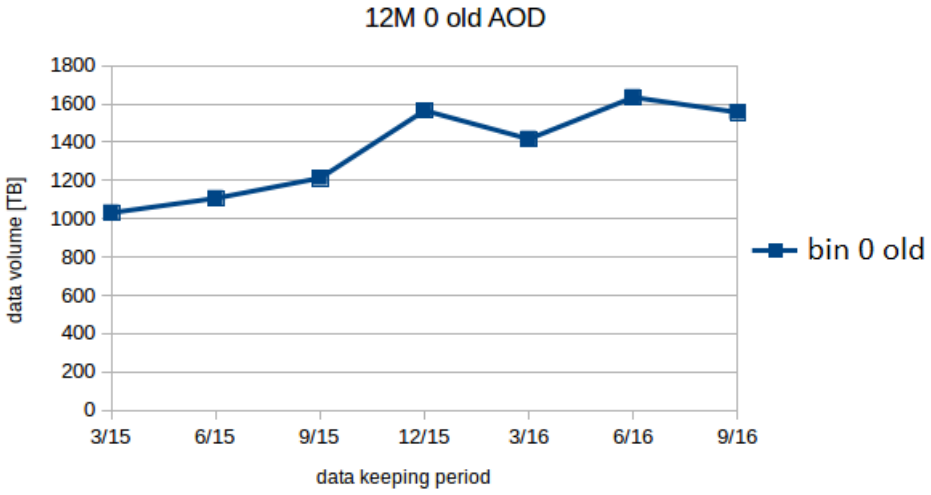

Figure D.3:

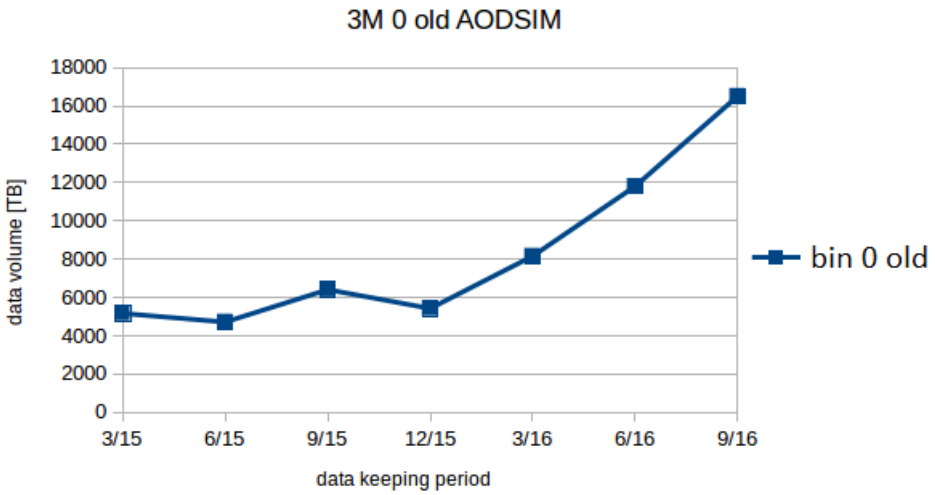

Figure D.4:

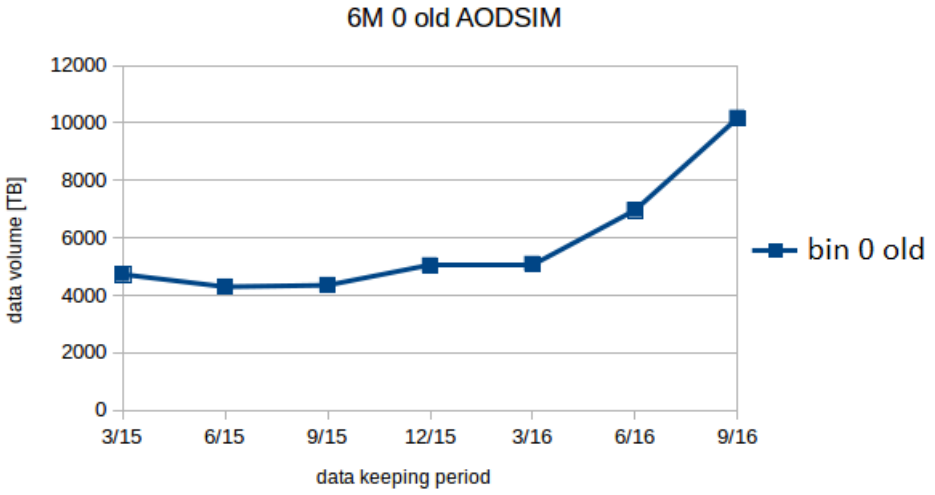

Figure D.5:

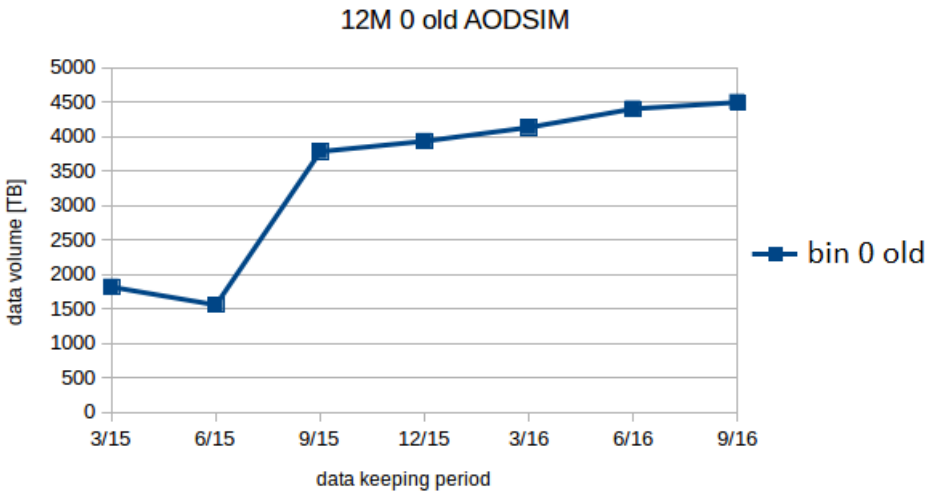

Figure D.6:

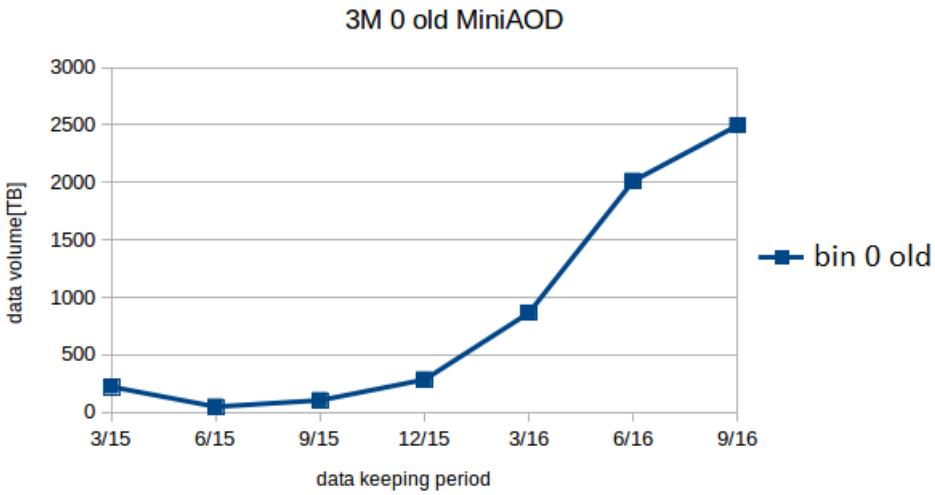

Figure D.7:

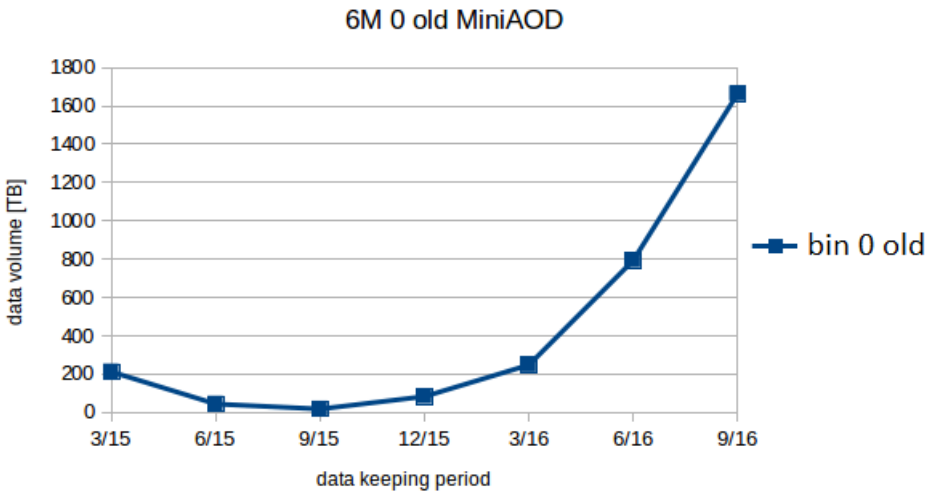

Figure D.8:

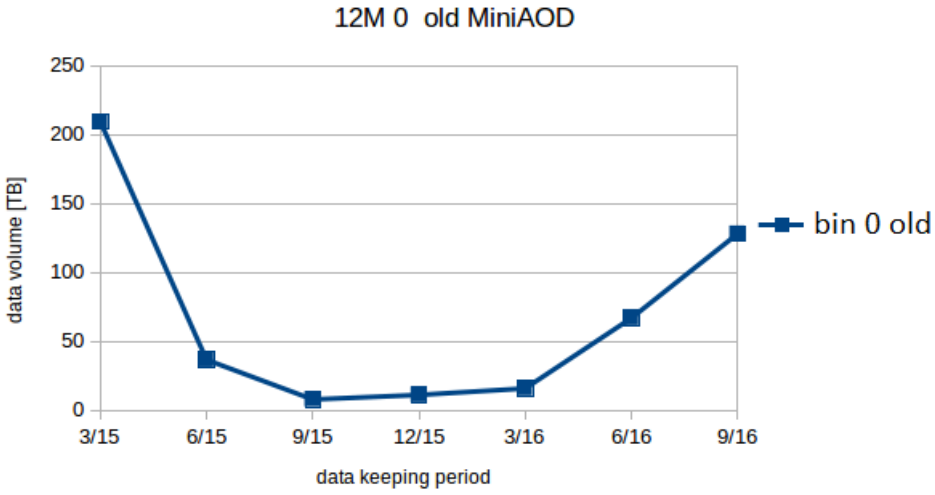

Figure D.9:

Appendix E

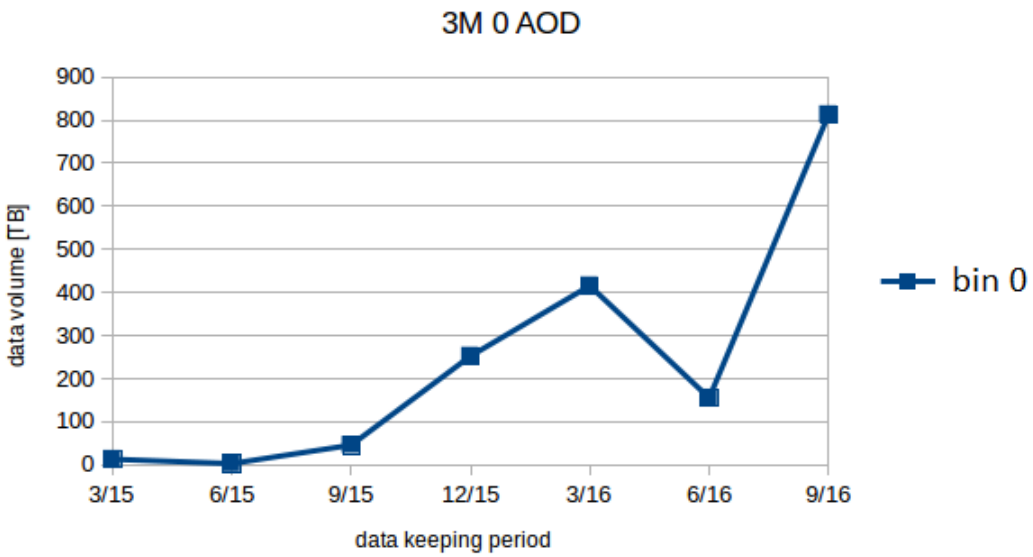

Figure E.1:

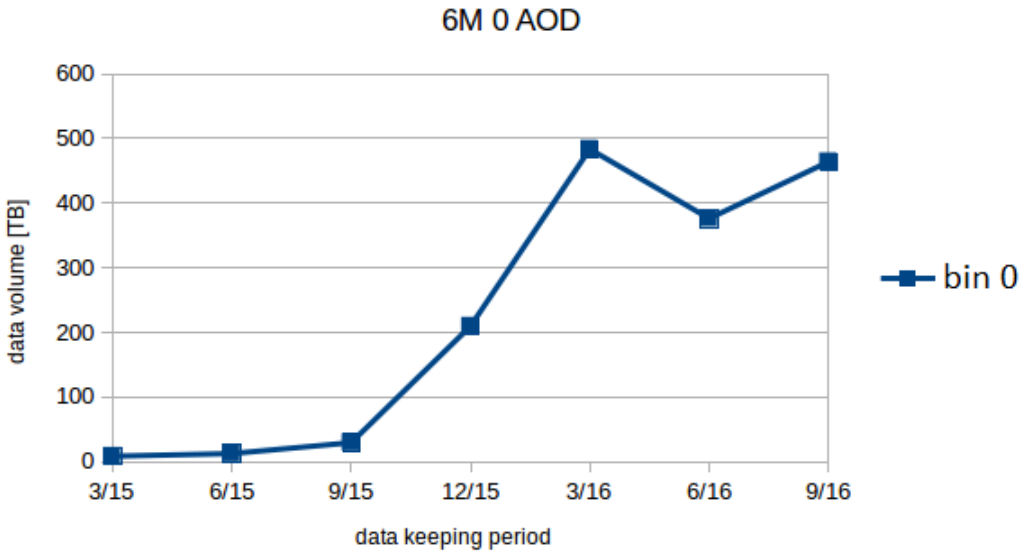

Figure E.2:

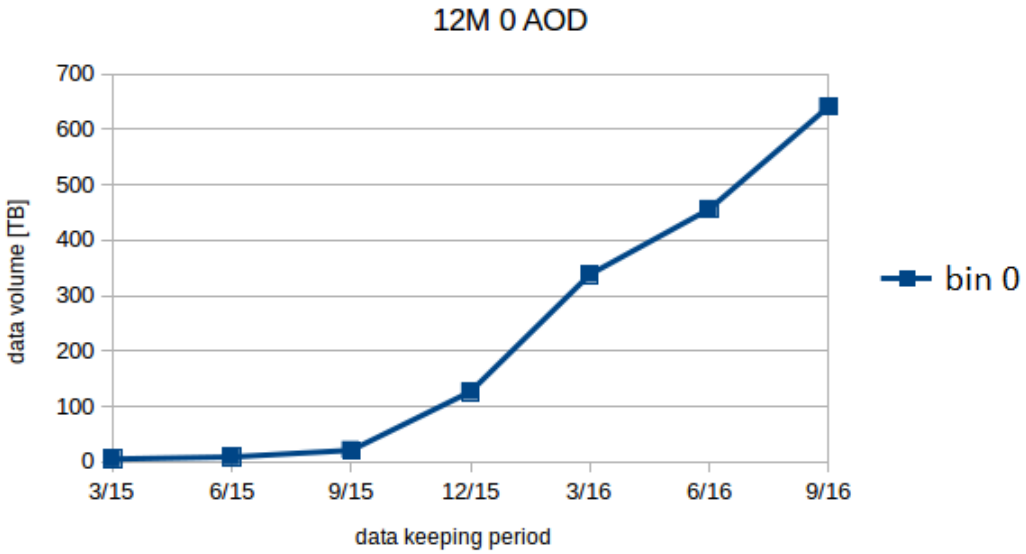

Figure E.3:

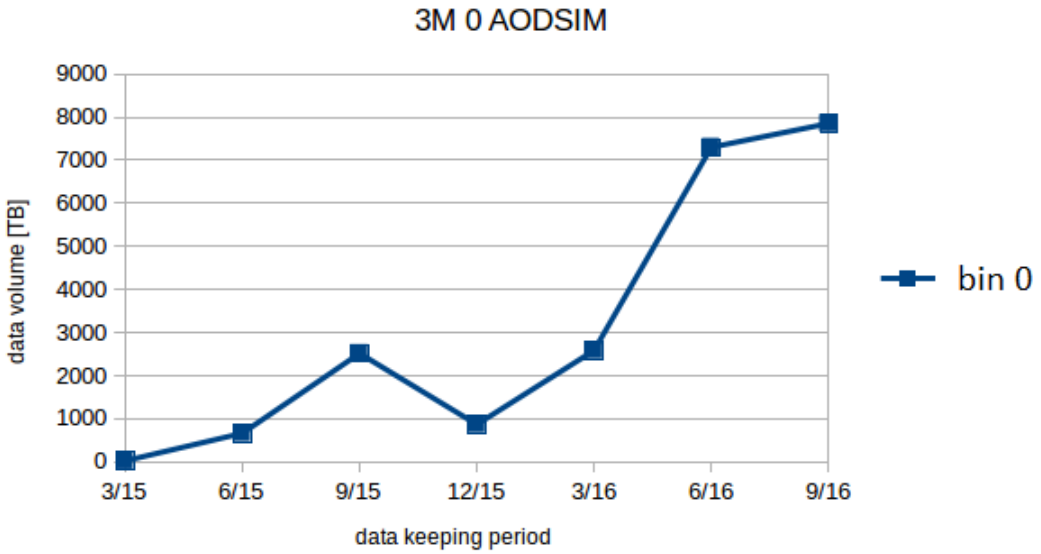

Figure E.4:

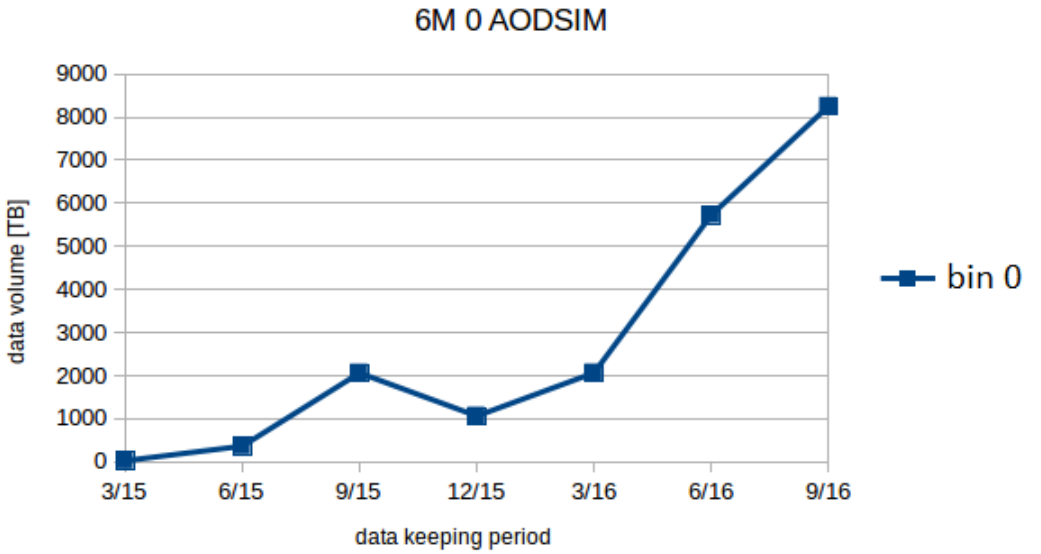

Figure E.5:

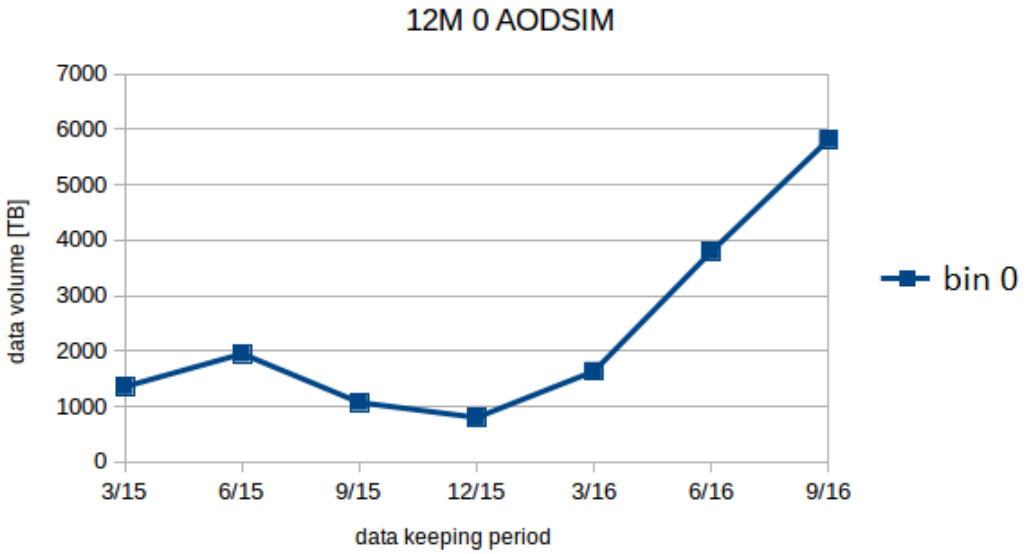

Figure E.6:

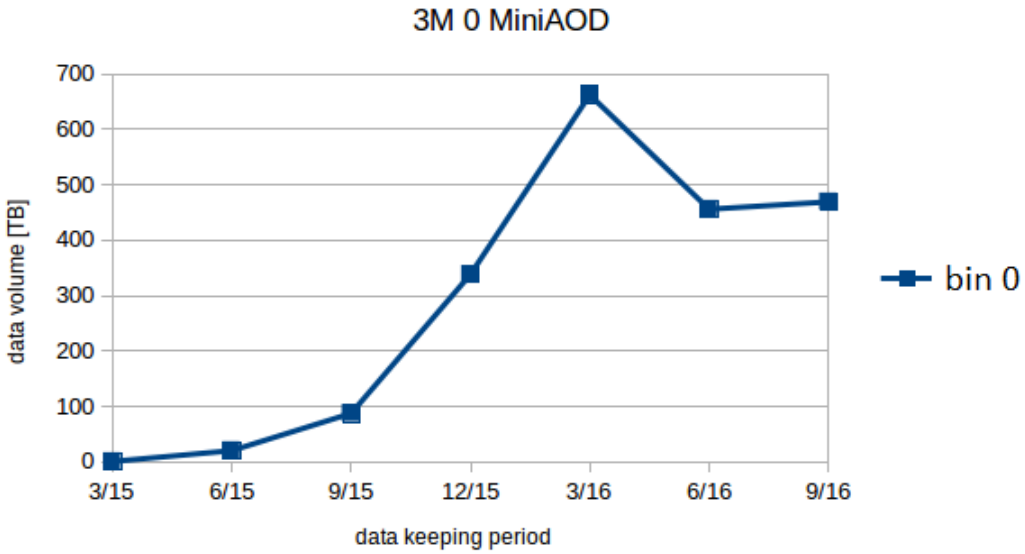

Figure E.7:

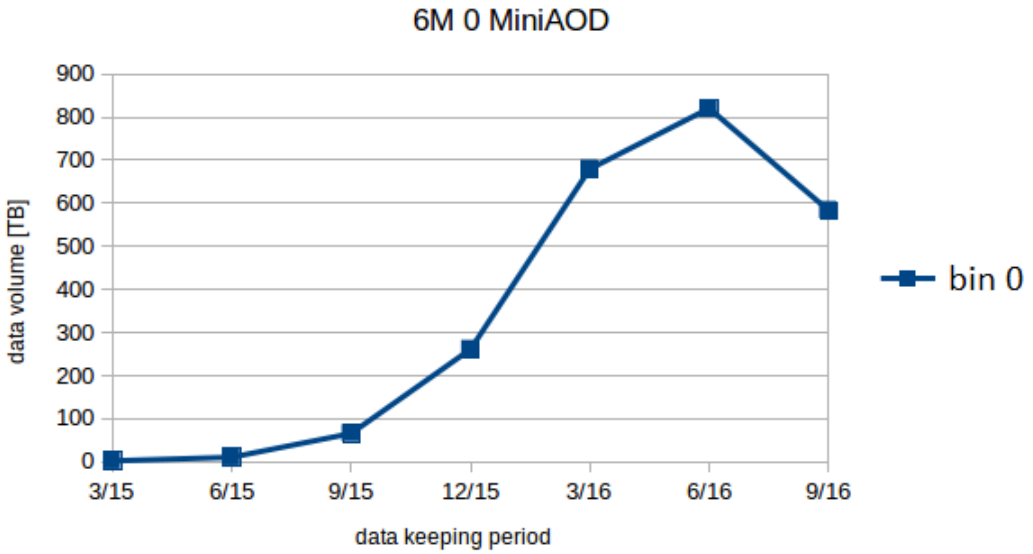

Figure E.8:

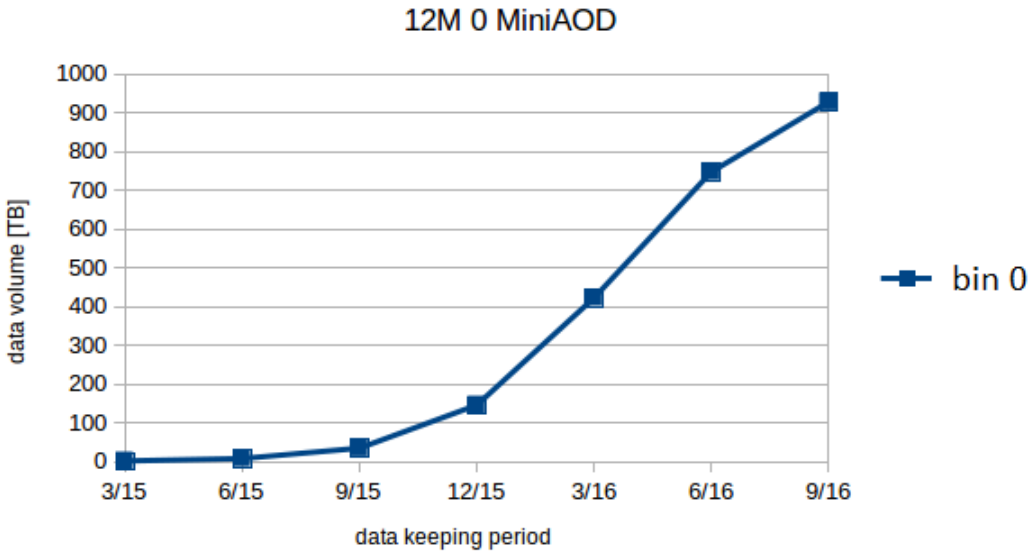

Figure E.9:

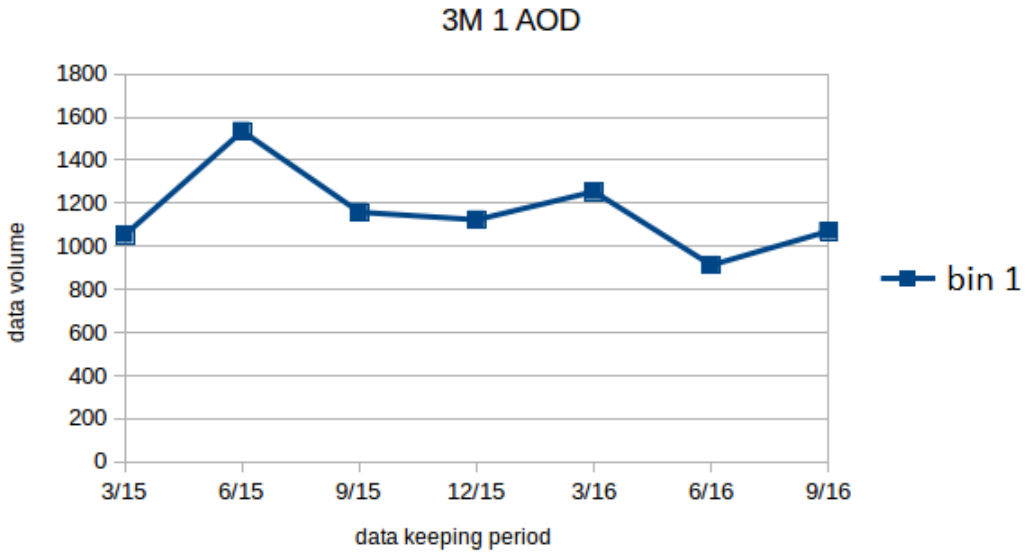

Figure E.10:

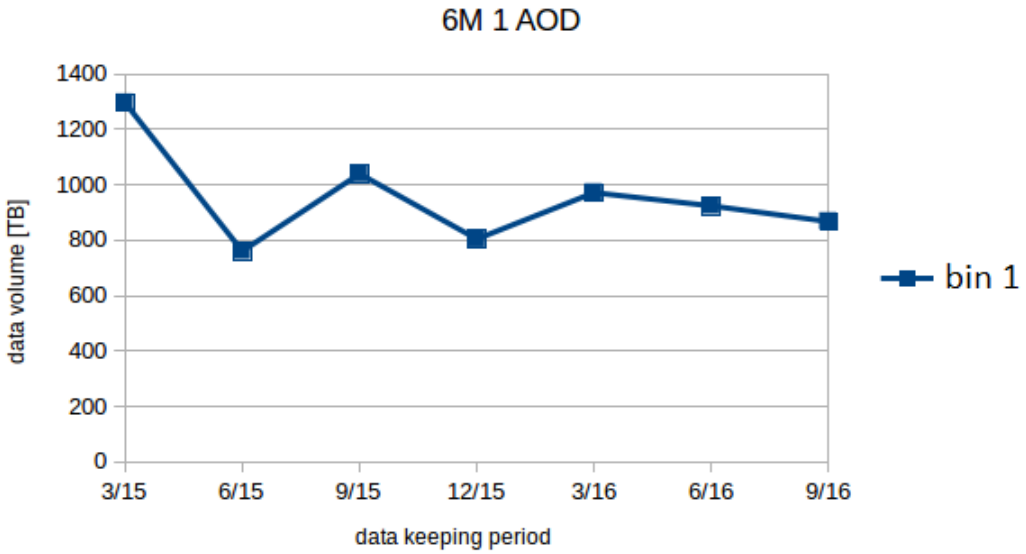

Figure E.11:

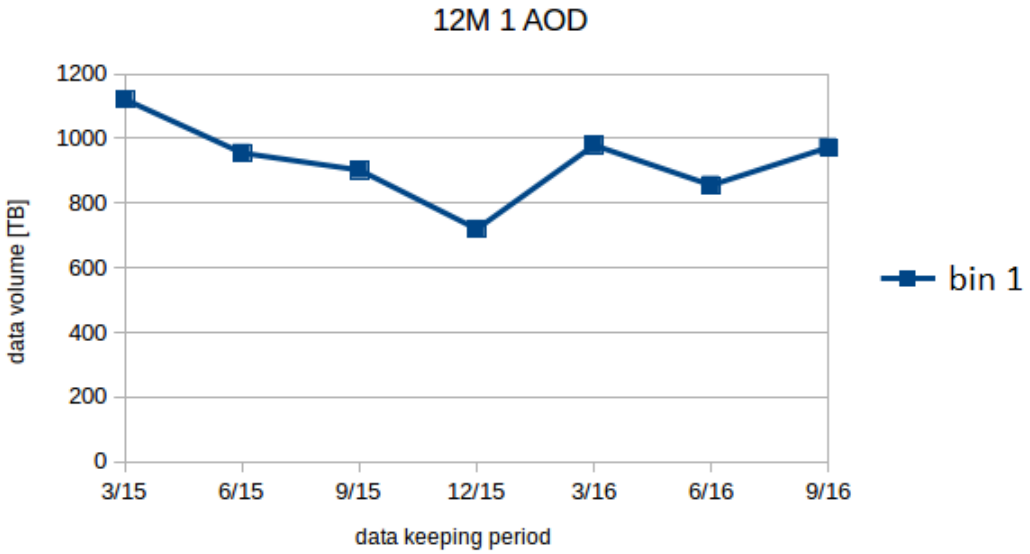

Figure E.12:

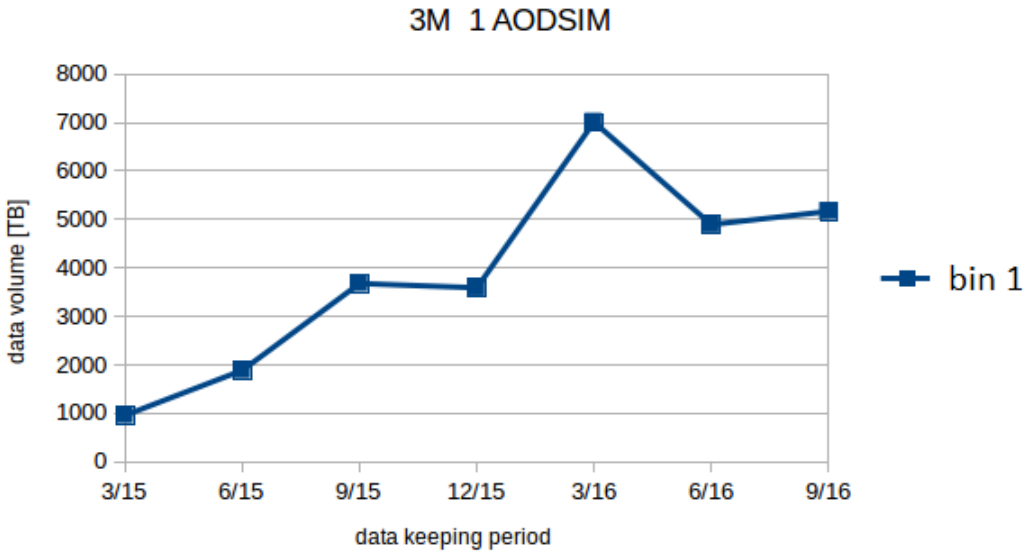

Figure E.13:

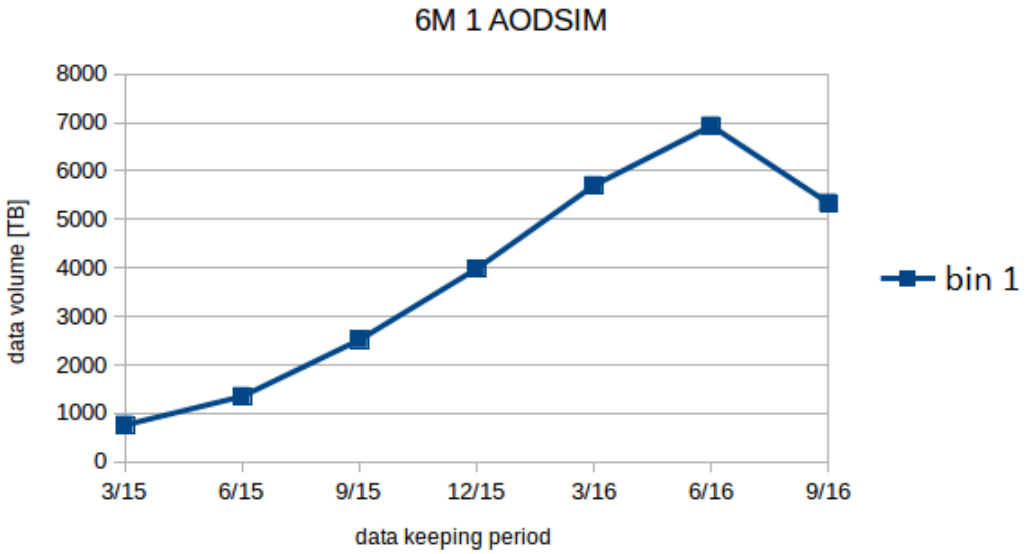

Figure E.14:

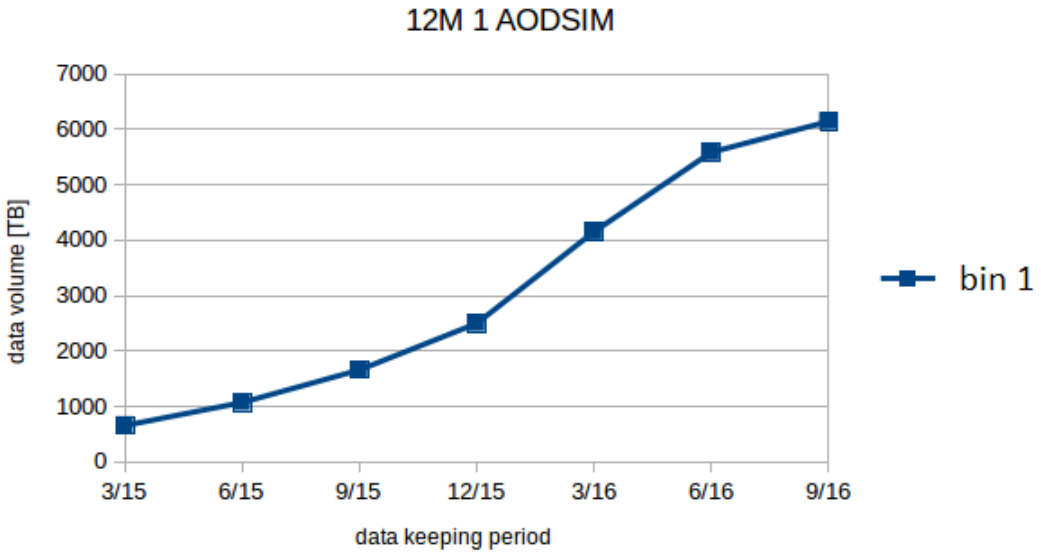

Figure E.15:

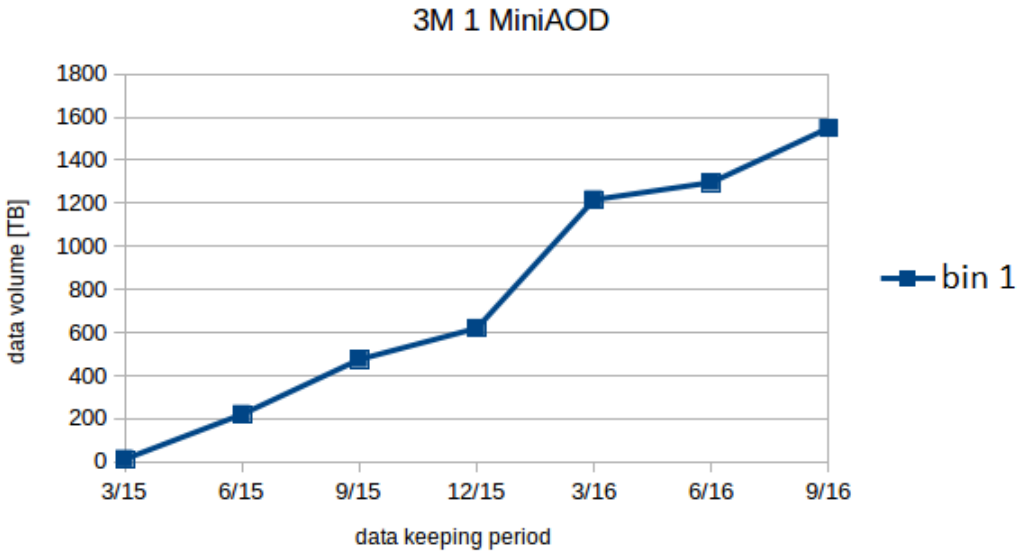

Figure E.16:

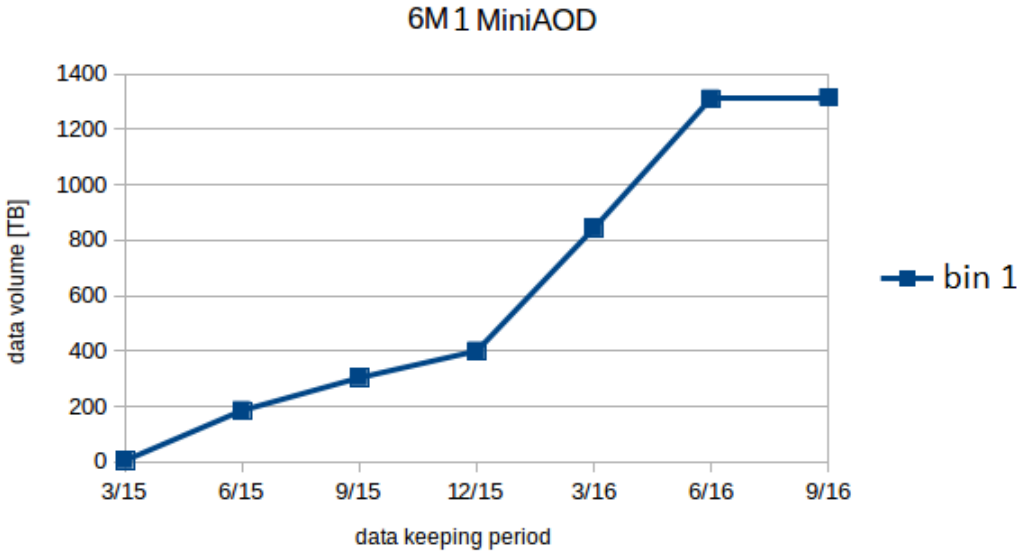

Figure E.17:

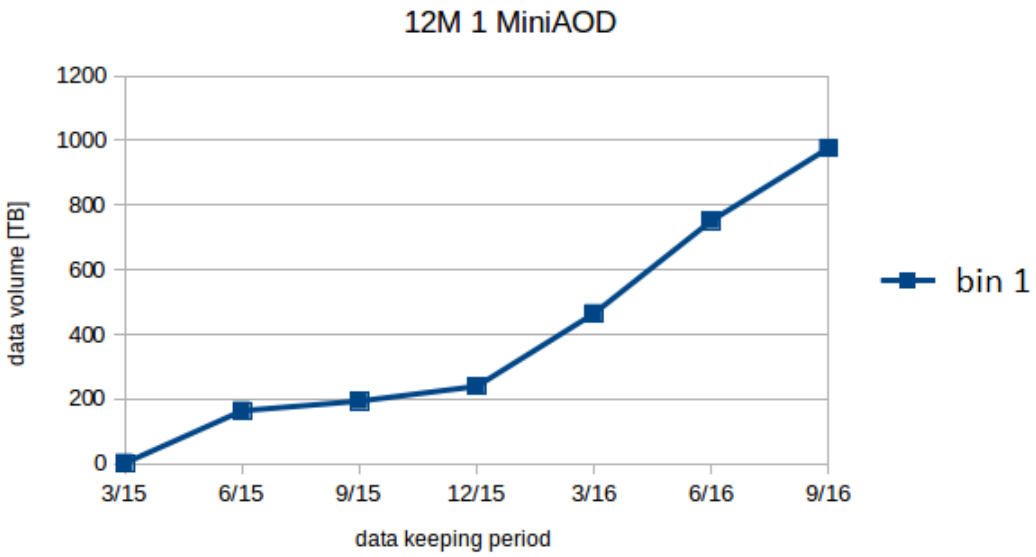
12M 1 MiniAOD
1200
1000
800
600
400
200
0
data volume [TB]
3/15
6/15
9/15
12/15
3/16
6/16
9/16
data keeping period
bin 1

Figure E.18:

Appendix F

Figure F.1:

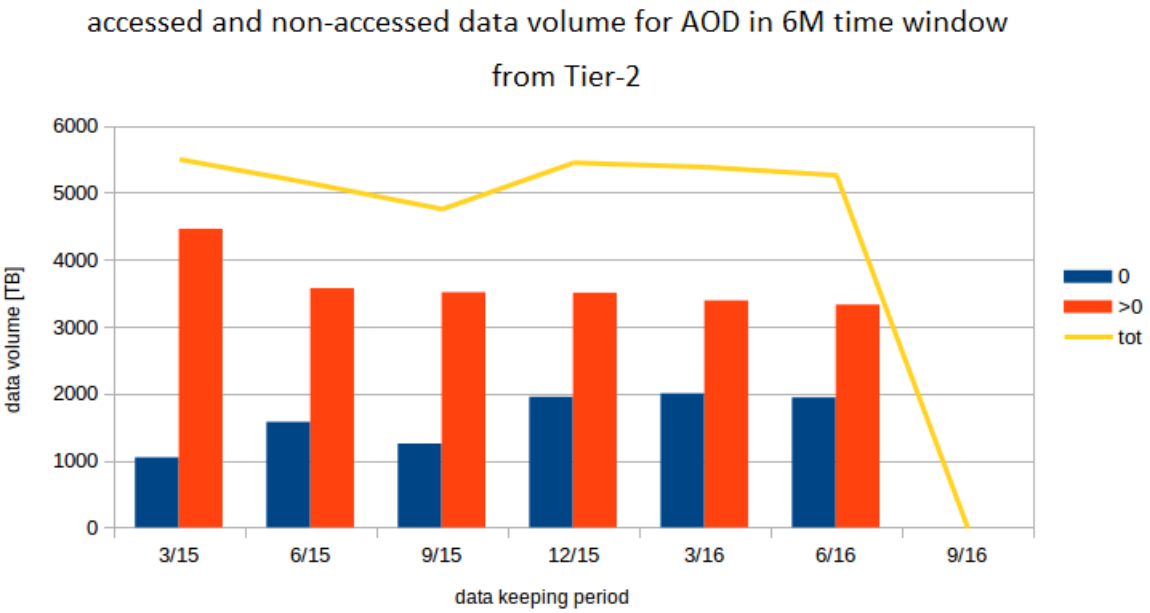

Figure F.2:

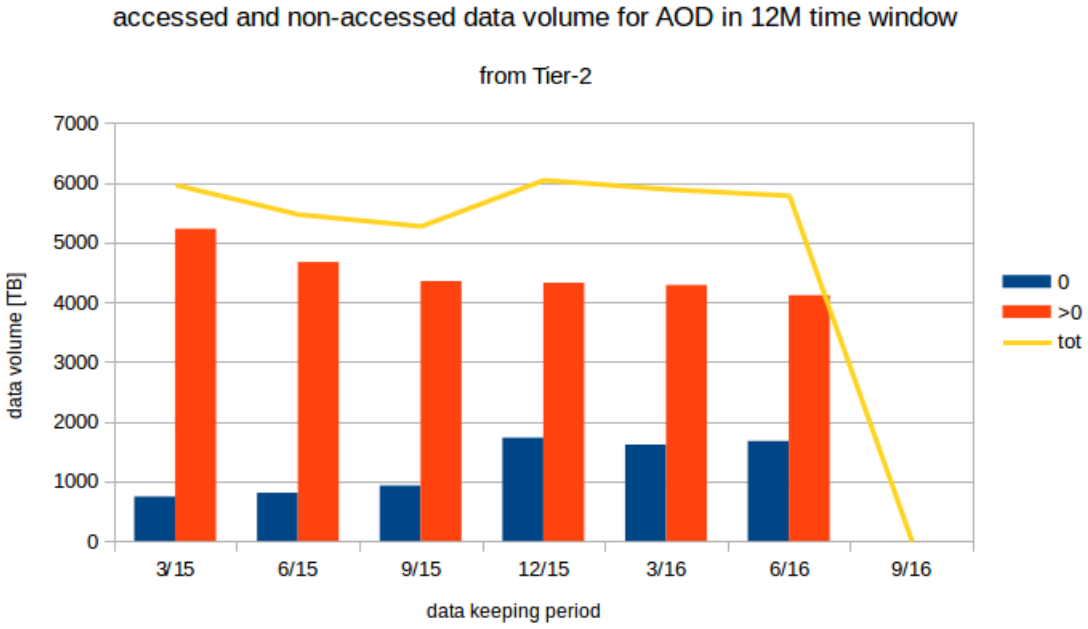

Figure F.3:

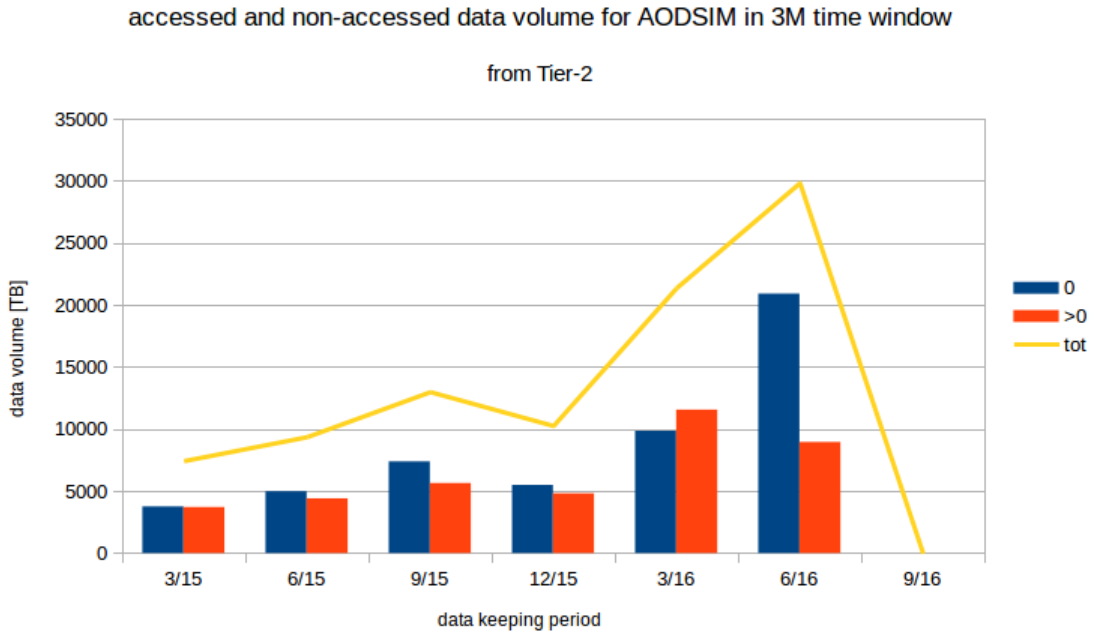

Figure F.4:

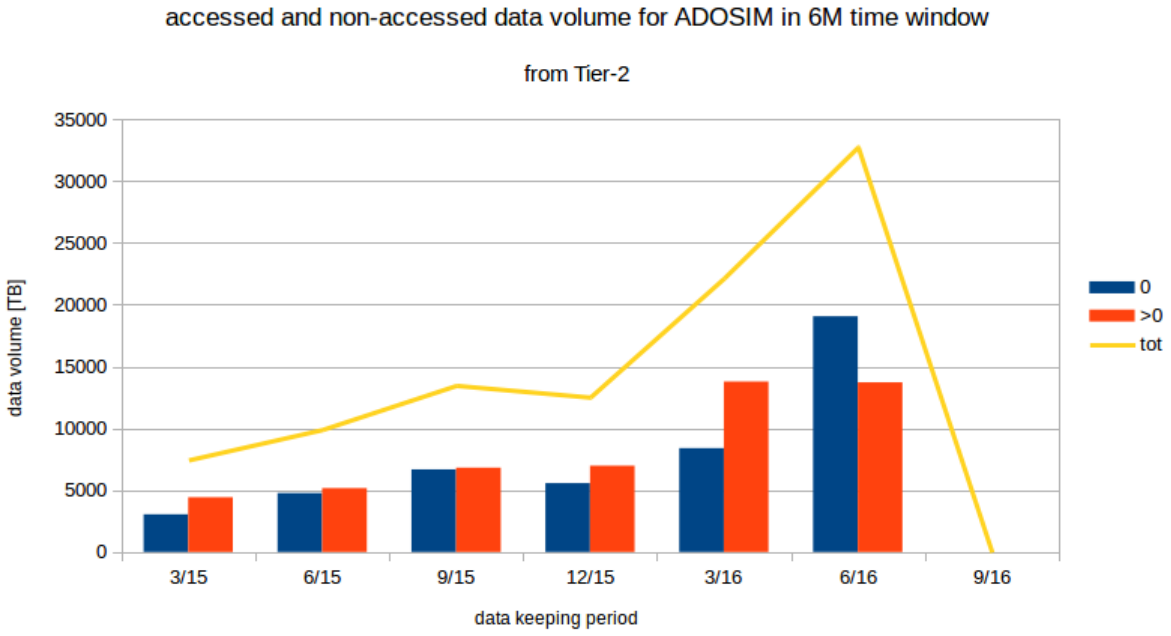

Figure F.5:

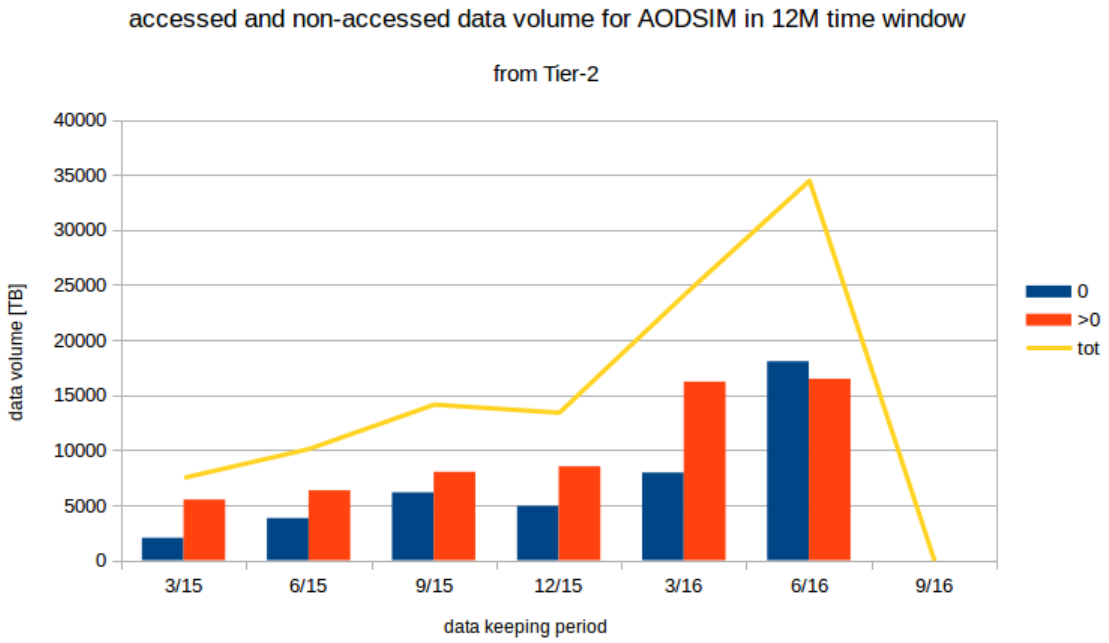

Figure F.6:

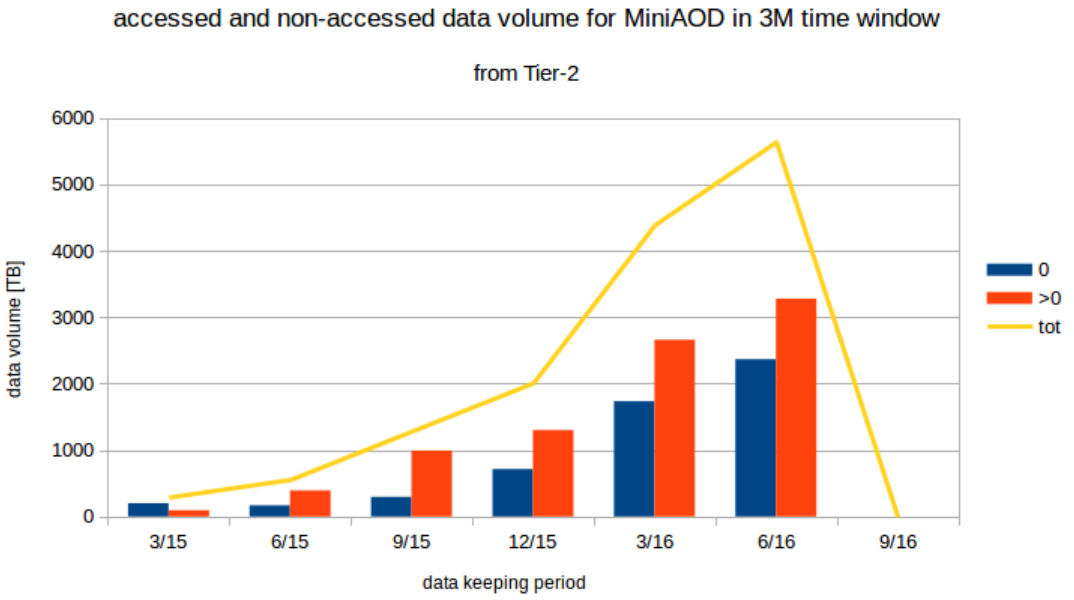

Figure F.7:

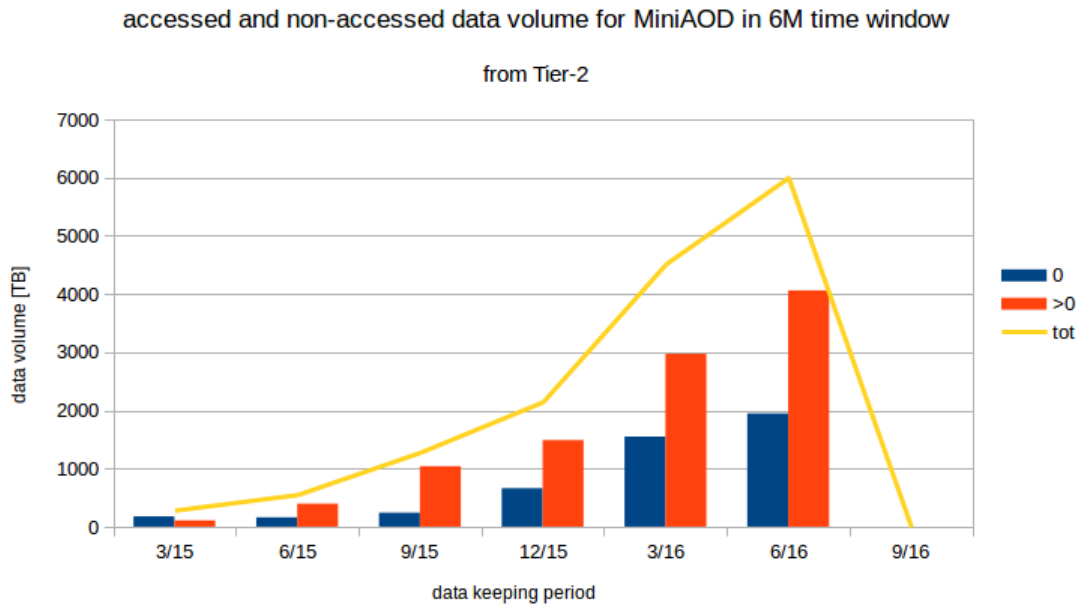

Figure F.8:

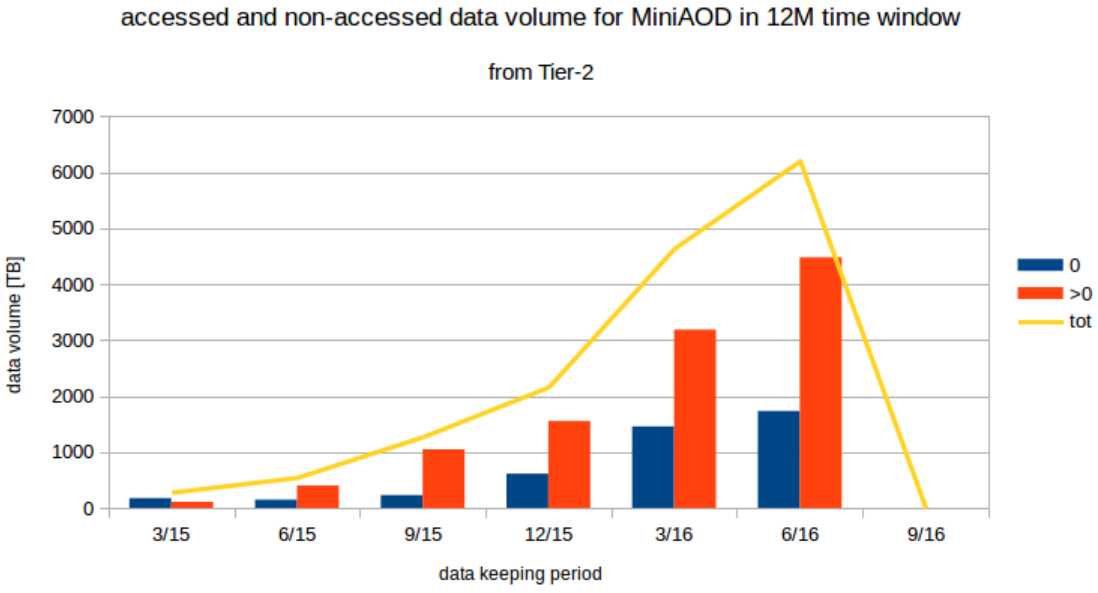

Figure F.9:

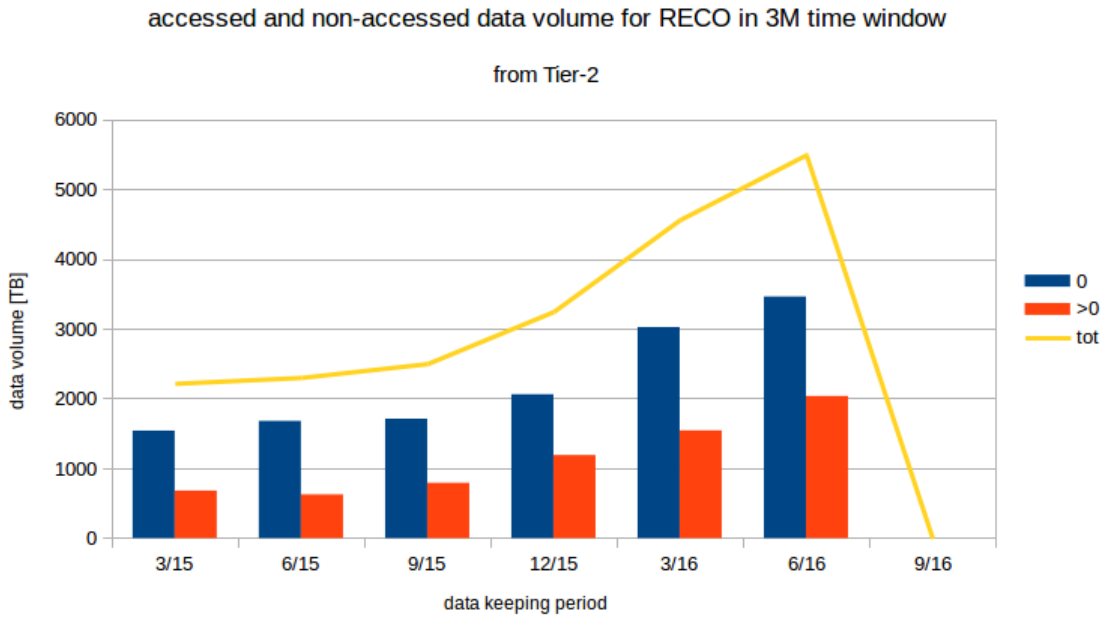

Figure F.10:

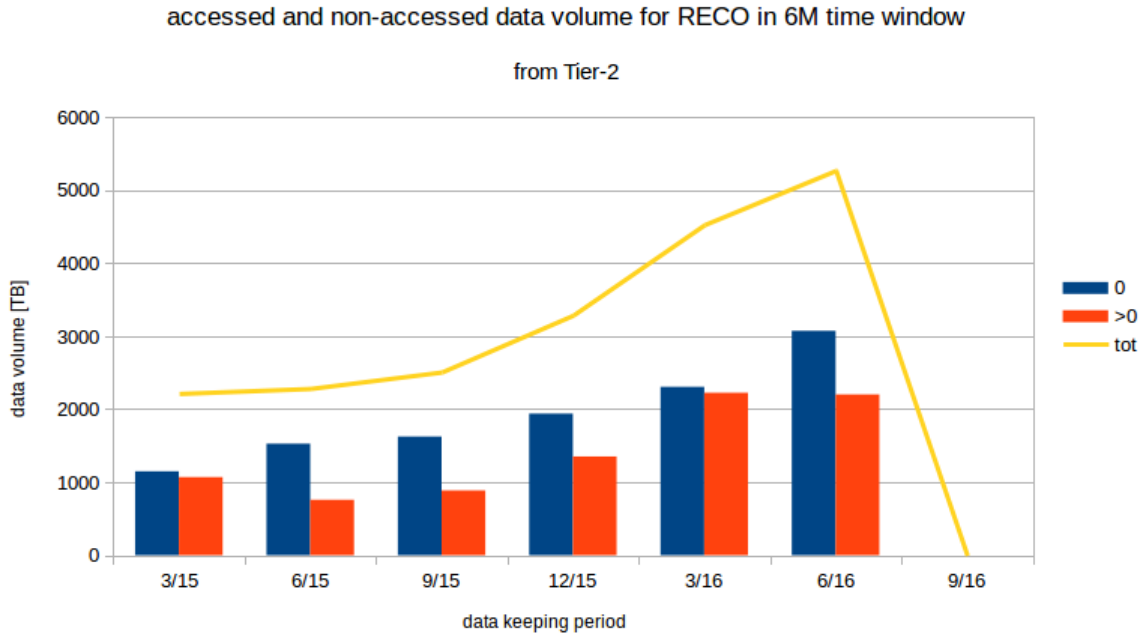

Figure F.11:

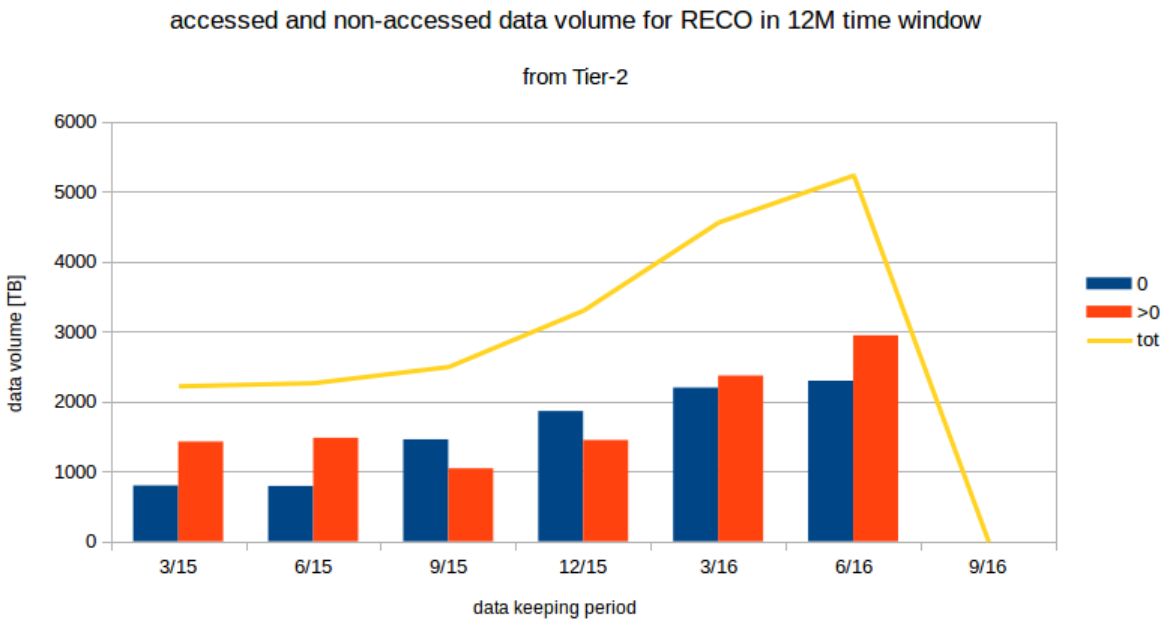
accessed and non-accessed data volume for RECO in 12M time window
from Tier-2
data volume [TB]
6000
5000
4000
3000
2000
1000
0
3/15
6/15
9/15
12/15
3/16
6/16
9/16
data keeping period
0
>0
tot

Figure F.12:

Appendix G

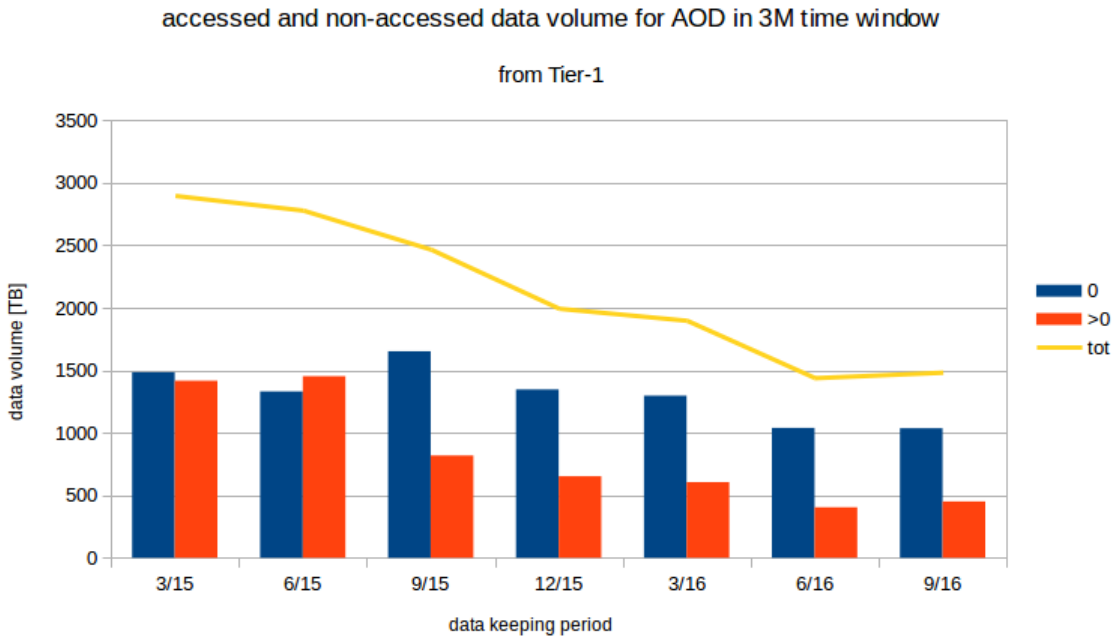

Figure G.1:

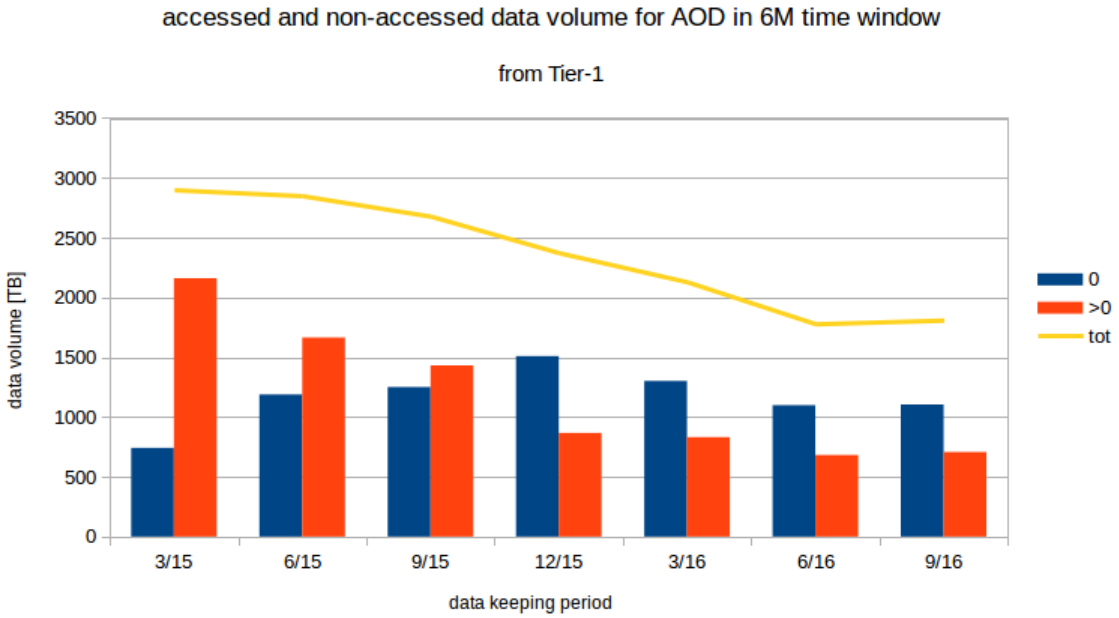

Figure G.2:

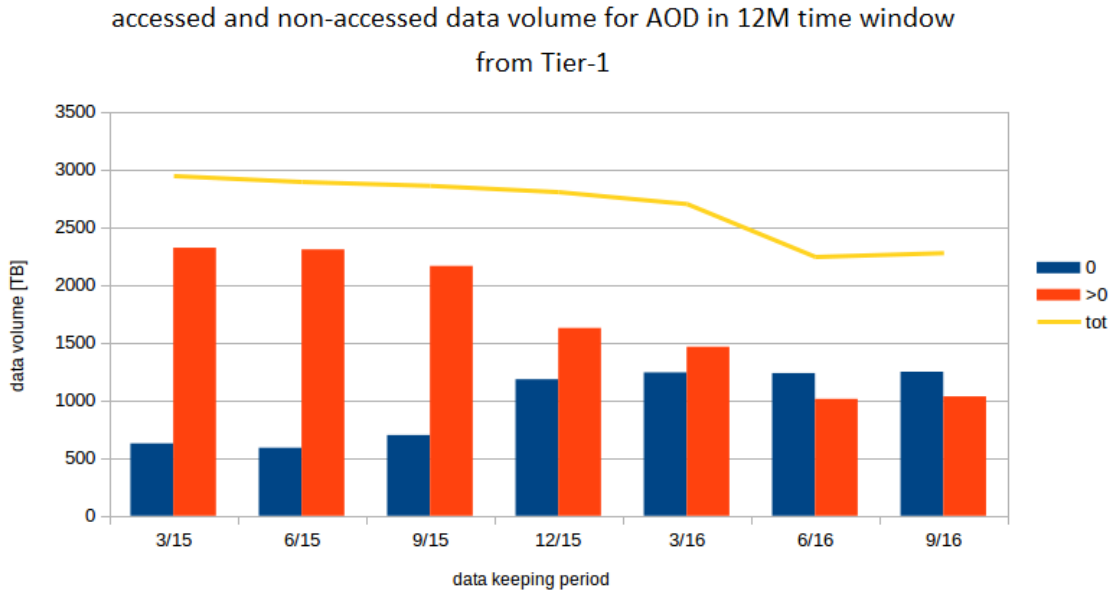

Figure G.3:

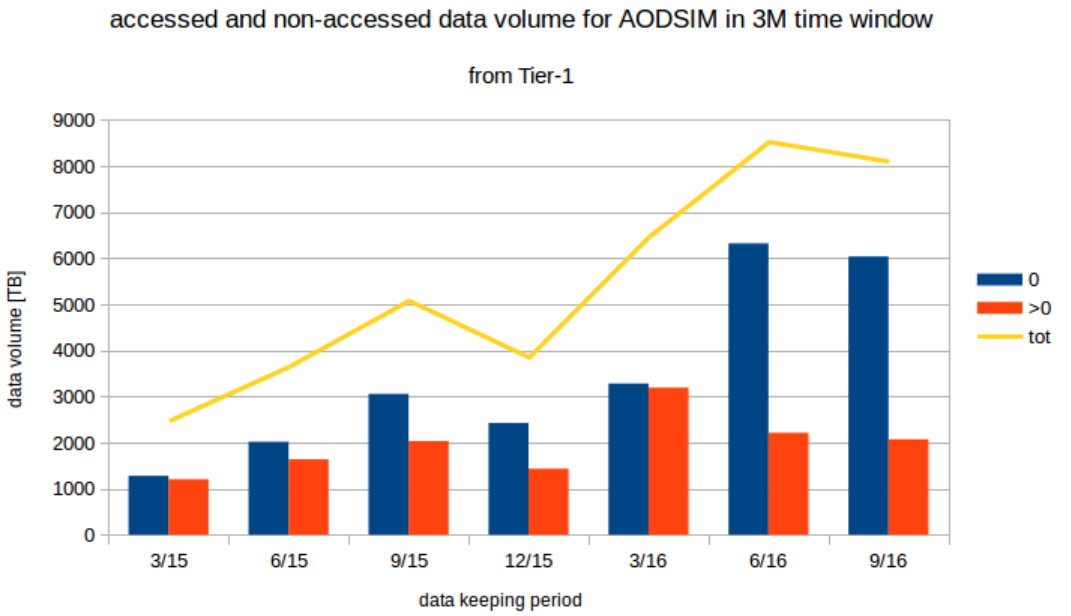

Figure G.4:

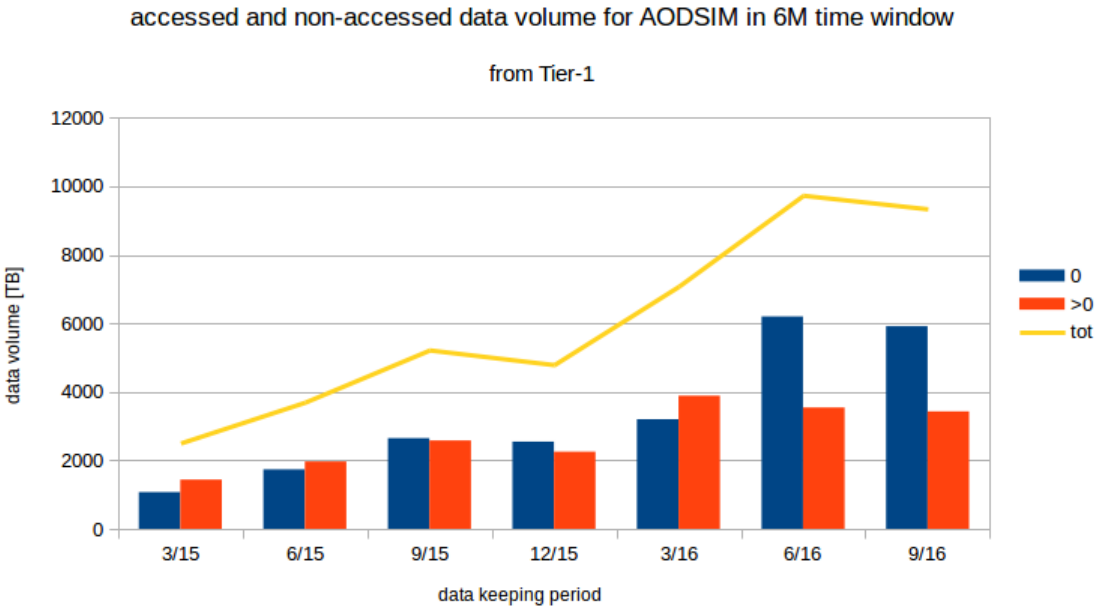

Figure G.5:

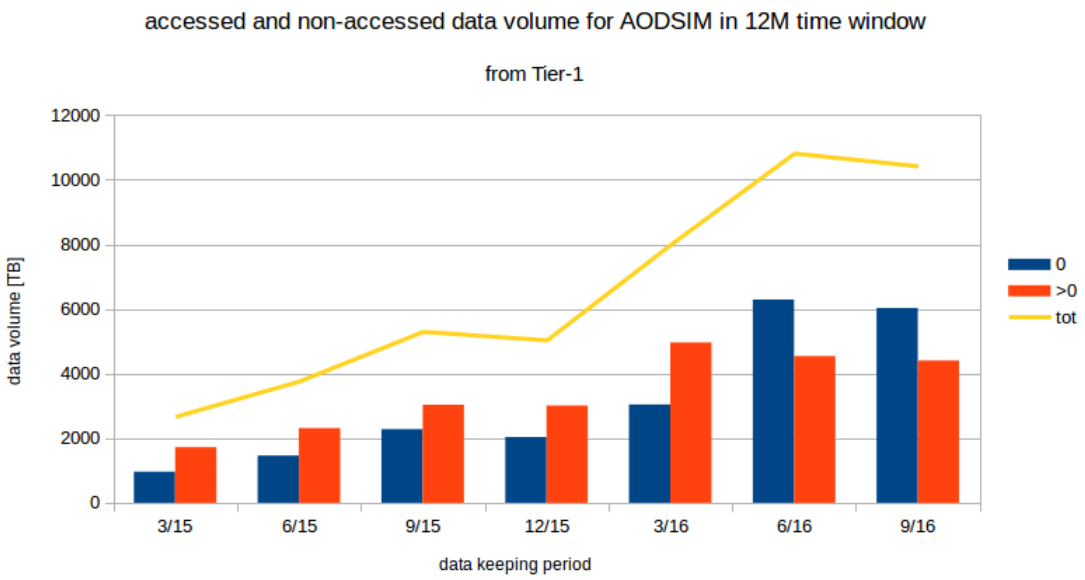

Figure G.6:

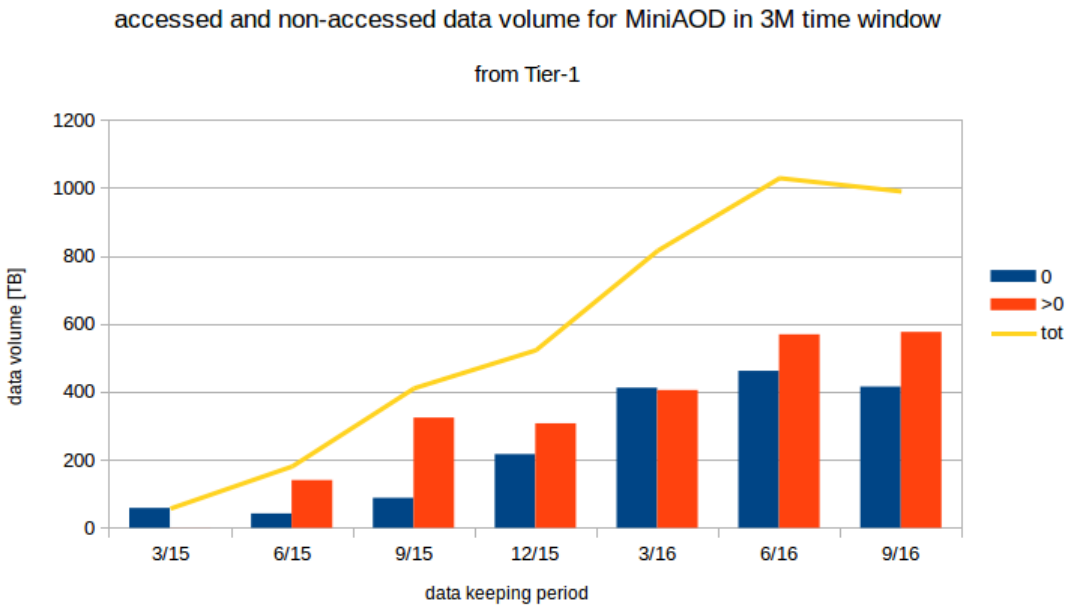

Figure G.7:

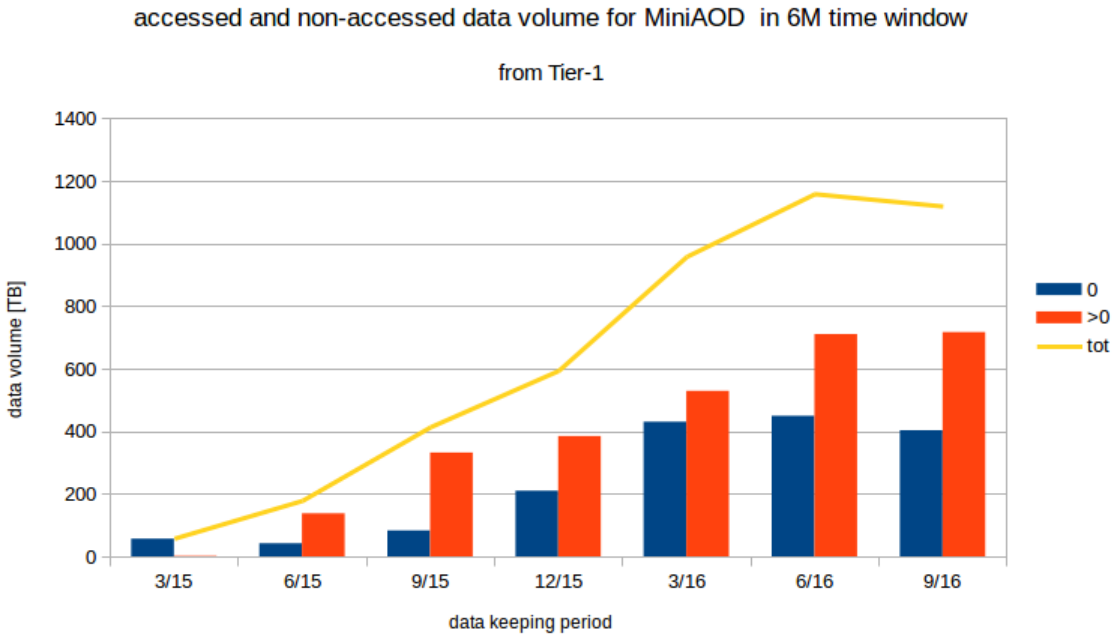

Figure G.8:

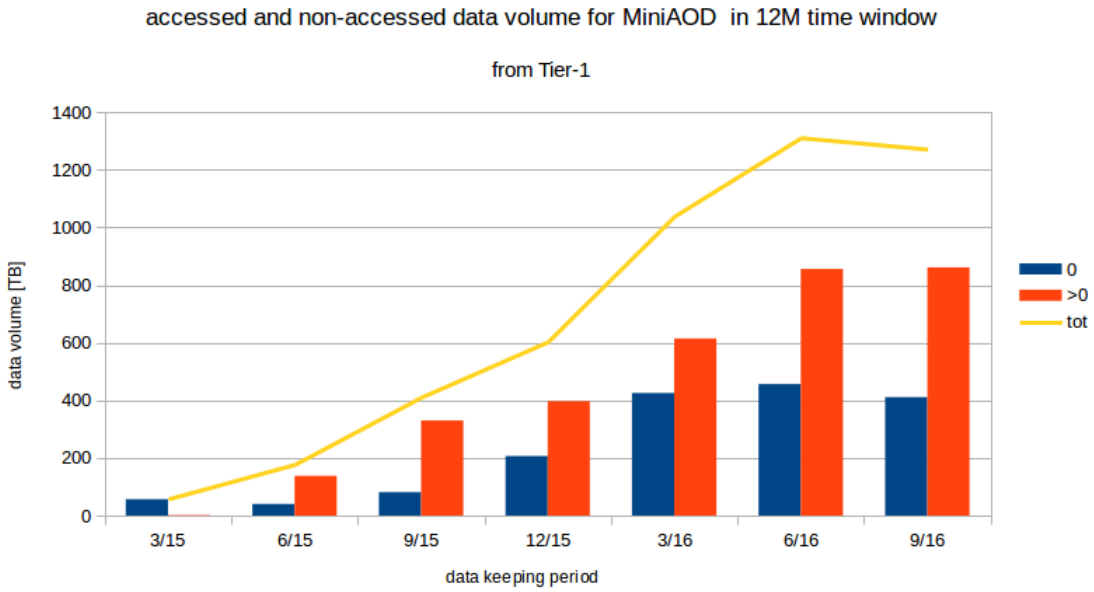

Figure G.9:

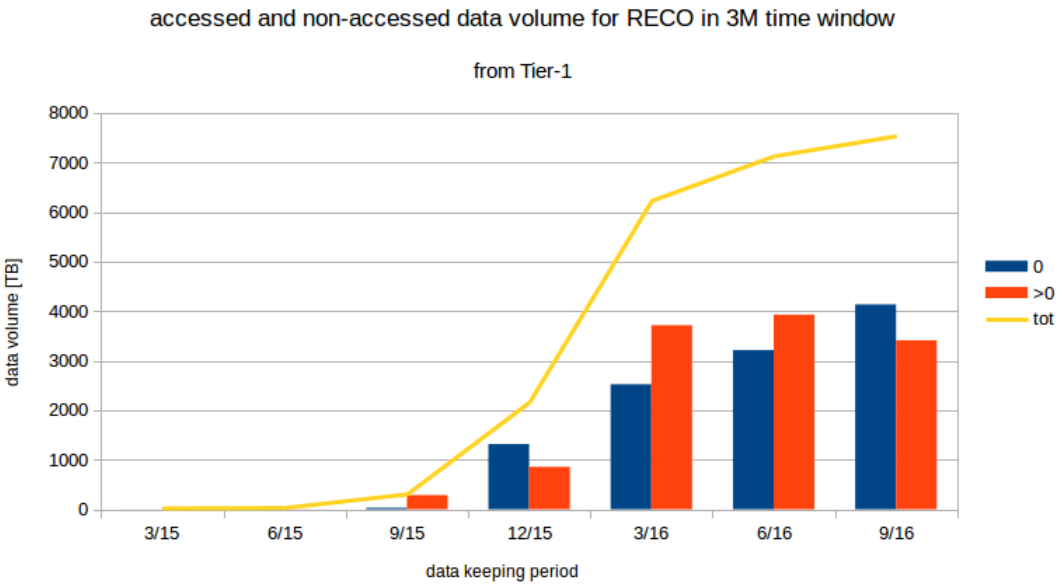

Figure G.10:

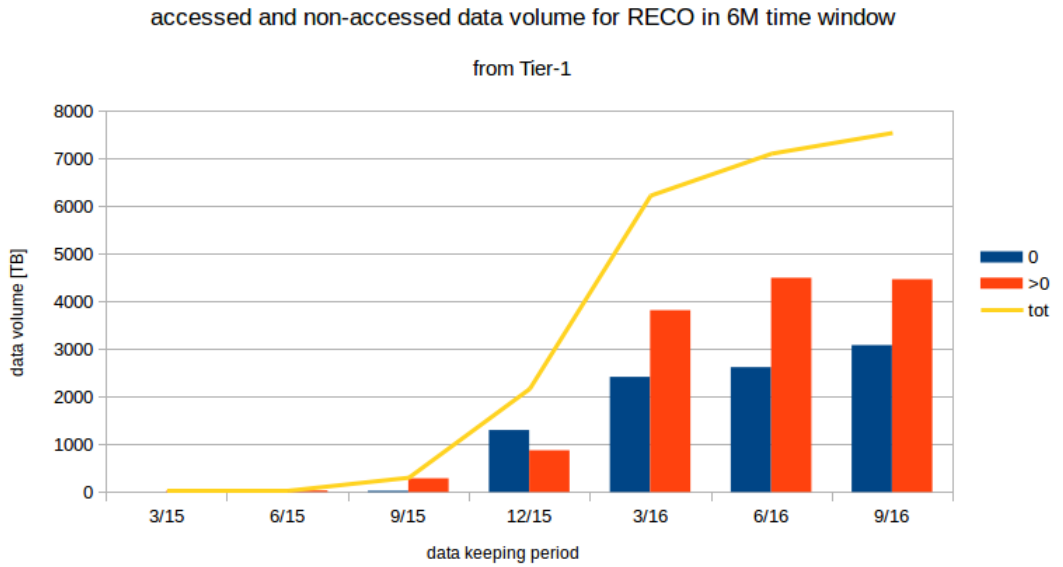

Figure G.11:

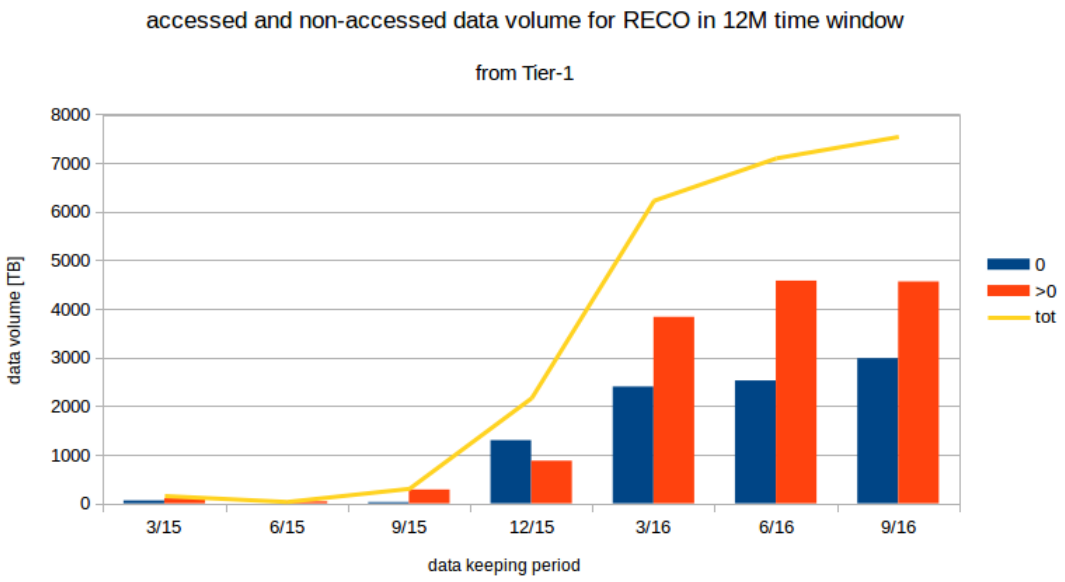
accessed and non-accessed data volume for RECO in 12M time window
from Tier-1
data volume [TB]
8000
7000
6000
5000
4000
3000
2000
1000
0
3/15
6/15
9/15
12/15
3/16
6/16
9/16
data keeping period
0
>0
tot

Figure G.12:

Appendix H

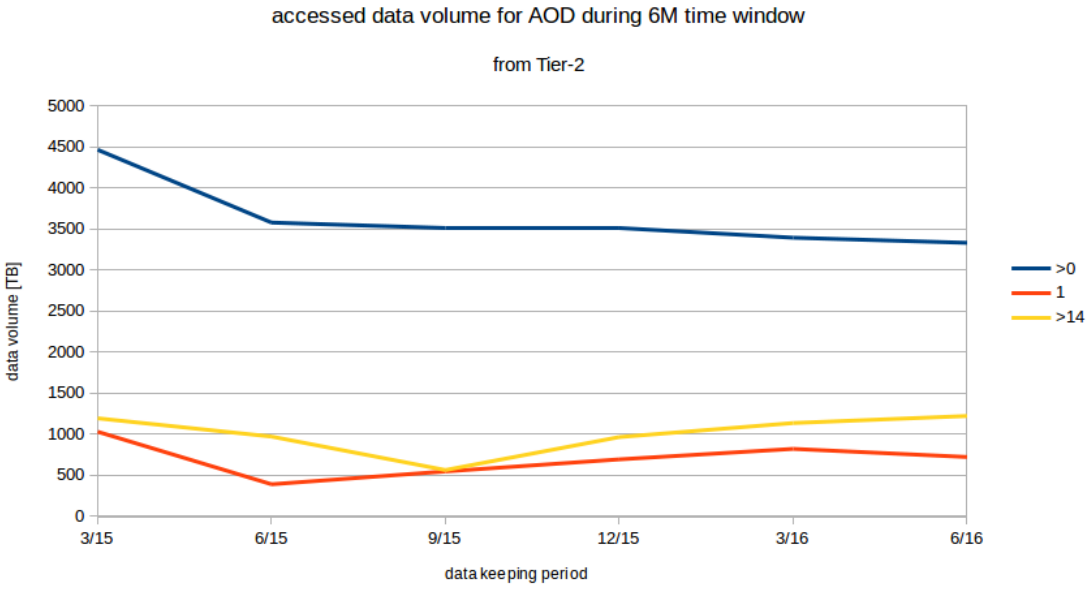

Figure H.1:

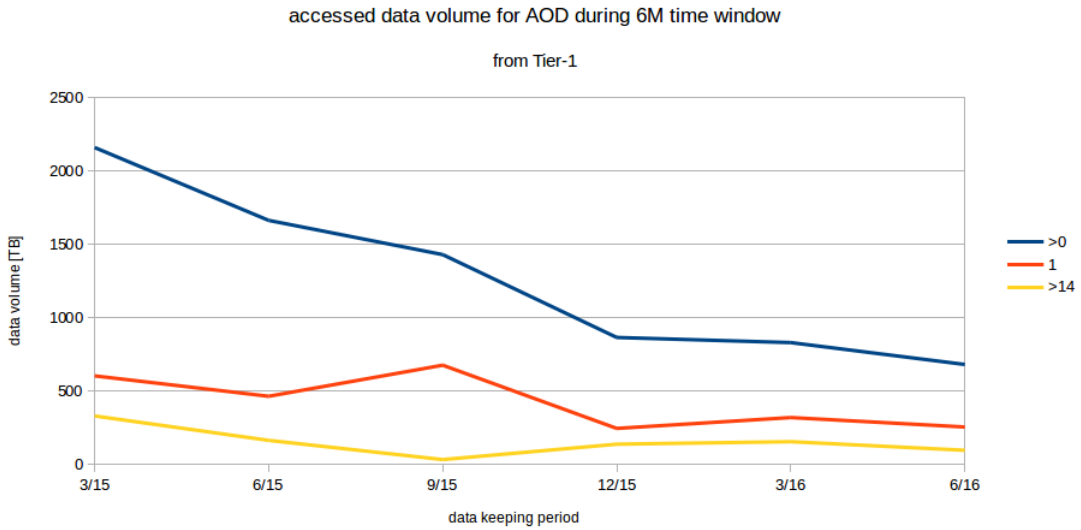

Figure H.2:

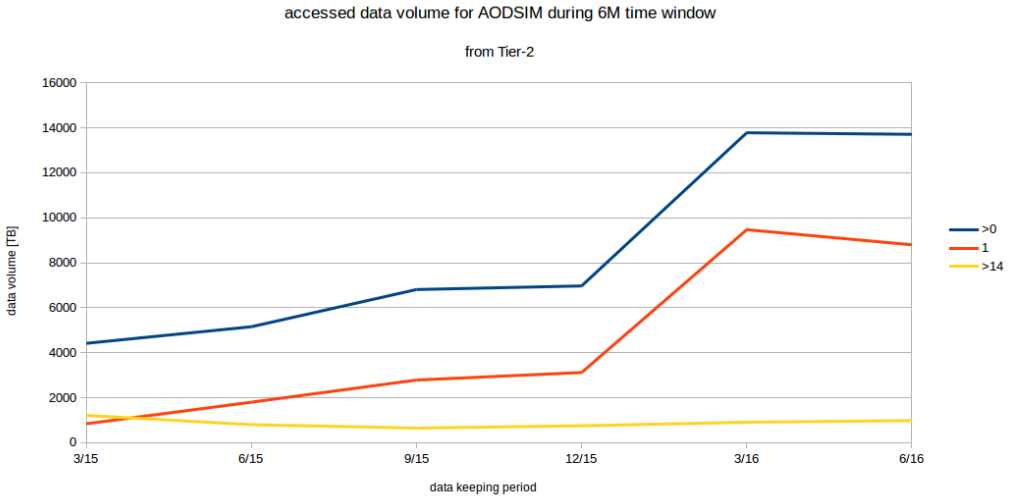

Figure H.3:

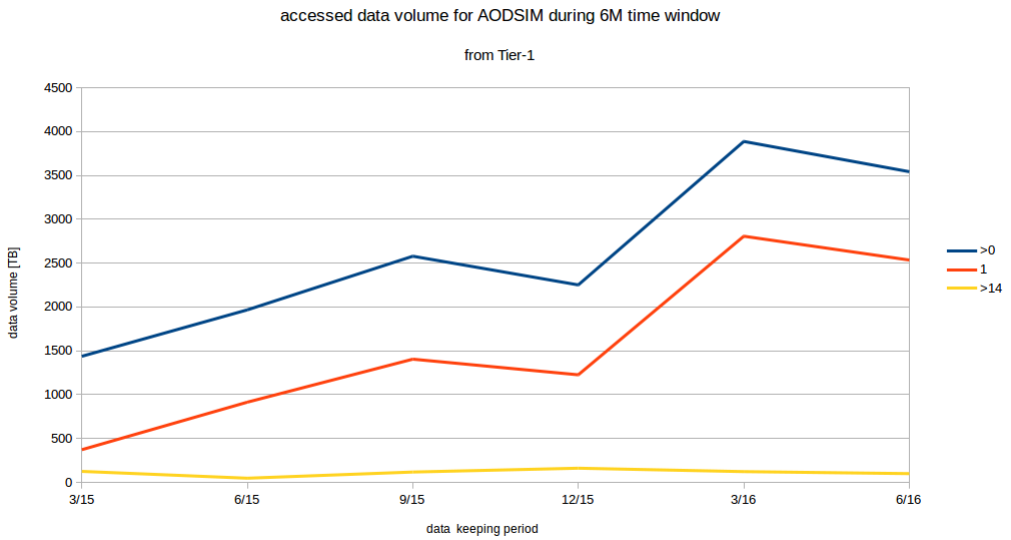

Figure H.4:

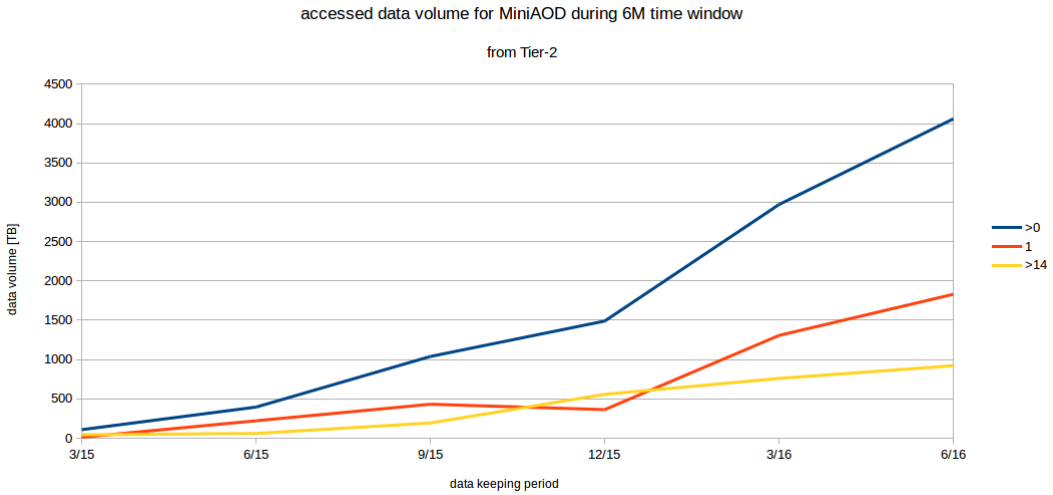

Figure H.5:

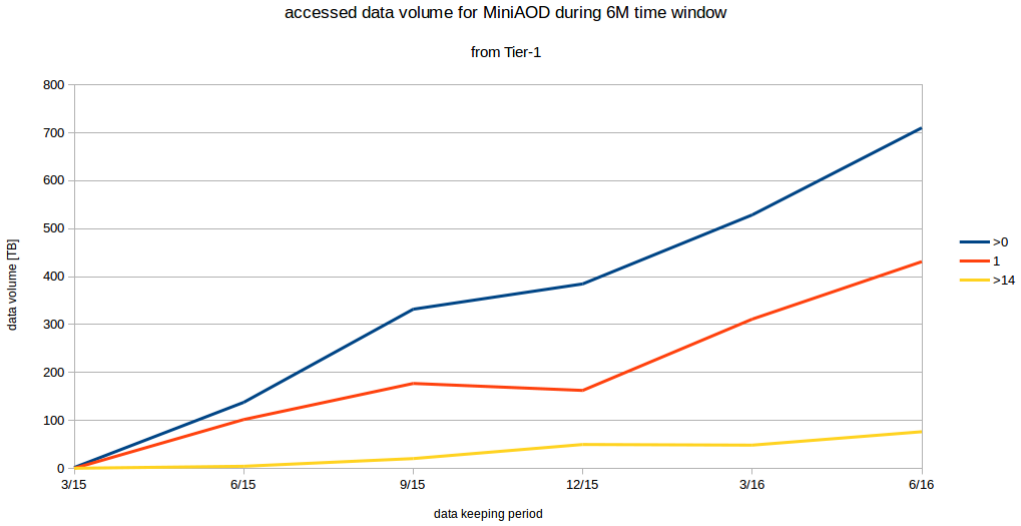

Figure H.6:

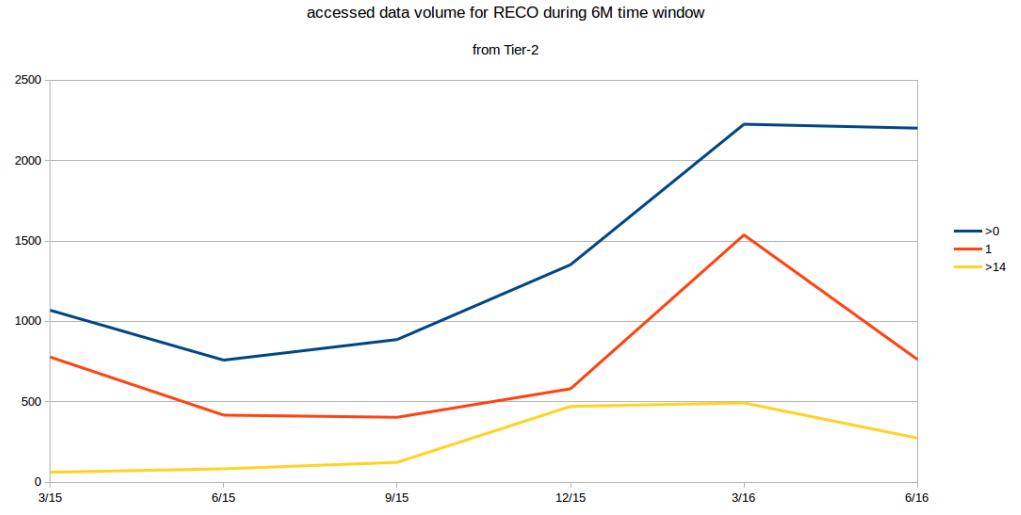

Figure H.7:

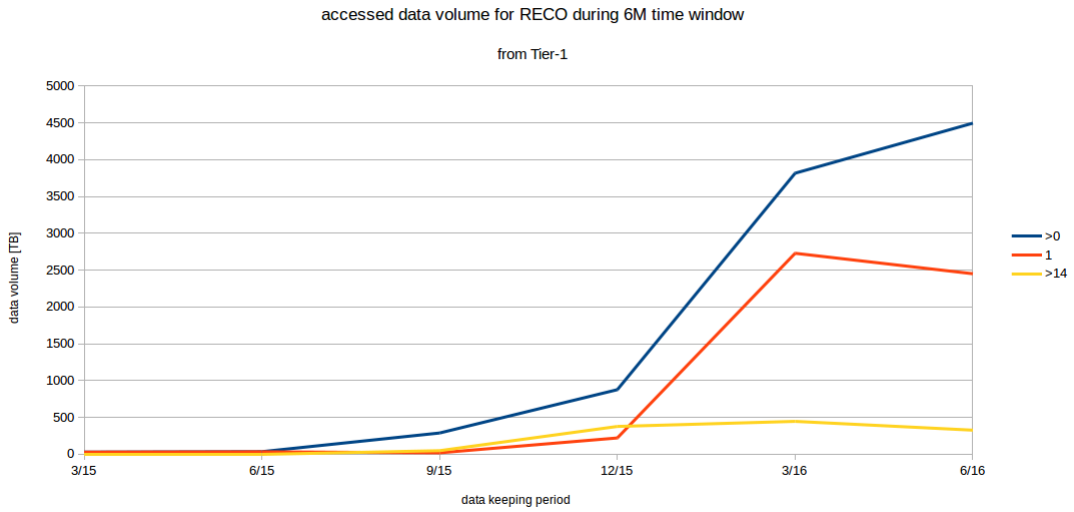
accessed data volume for RECO during 6M time window
from Tier-1
data volume [TB]
5000
4500
4000
3500
3000
2500
2000
1500
1000
500
0
3/15
6/15
9/15
12/15
3/16
6/16
data keeping period
>0
1
>14

Figure H.8:

Bibliography

[1]*About CERN.* URL:https://home.cern/about

[2] J. Krige. "History of CERN Vol.3", 1st Edition (1996).

[3] W.N. Cottingham, D. A. Greenwood. "An Introduction to the Standard Model of Particle Physics", Ed. by CAMBRIDGE, 2007.

[4]K.A. Olive et al. "Particle Data Group", Chin. Phys. C, 38,090001 (2014).

[5] Oliver Sim Brüning et al. *LHC Design Report.* Ed. by CERN library copies. Vol. 1, 2 ,3. 2012. URL:http://ab-div.web.cern.ch/ab-div/Publications/LHC-DesignReport.html

[6] Maurizio Vretenar "Linear accelerators". In: *CERN Yellow Report*,CERN-2013-001 (2013). URL:https://arxiv.org/ftp/arxiv/papers/1303/1303.6766.pdf

[7] Lyndon Evans and Philip Bryant. "LHC Machine". In: *Journal of Instrumentation* 3.08 (2008). Ed. by IOPscience, S08001. http://iopscience.iop.org/1748-0221/3/08/S08001

[8] J. M. Jimenez. "The CERN LHC - World's largest vacuum systems". In: WE4RAI02 (2009) Ed. by Proceedings of PAC09, Vancouver, BC, Canada. URL: https://accelconf.web.cern.ch/accelconf/PAC2009/papers/we4rai02.pdf

[9] *The Alice Collaboration.* URL: http://aliceinfo.cern.ch

[10] The ALICE Collaboration et al. "The ALICE experiment at the CERN LHC". In: *Journal of Instrumentation* 3.08 (2008). Ed. by IOPscience, S08002. URL:

[11] *The Atlas Collaboration.* URL:http://atlas.web.cern.ch/Atlas/Collaboration

[12] The ATLAS Collaboration et al. "The ATLAS Experiment at the CERN Large Hadron Collider". In: *Journal of Instrumentation* 3.08 (2008). Ed. by IOPscience, S08003. URL: http://iopscience.iop.org/1748-0221/3/08/S08003

[13] *The CMS Collaboration.* URL:http://cms.web.cern.ch

[14] *The LHCb Collaboration.* URL:http://lhcb.web.cern.ch/lhcb

[15] The LHCb Collaboration et al. "The LHCb Detector at the LHC". In: *Journal of Instrumentation* 3.08 (2008). Ed. by IOPscience, S08005. URL:http://iopscience.iop.org/1748-0221/3/08/S08005

[16] *The LHCf experiment.* URL:http://home.web.cern.ch/about/experiments/lhcf

[17] The LHCf Collaboration et al. "The LHCf detector at the CERN Large Hadron Collider". In: *Journal of Instrumentation* 3.S08006 (2008). Ed. by IOPscience. URL:http://iopscience.iop.org/1748-0221/3/08/S08006

[18] *The TOTEM Collaboration.* URL:http:/totem.web.cern.ch/Totem/

[19] The TOTEM Collaboration et al. "The TOTEM Experiment at the CERN Large Hadron Collider". In: *Journal of Instrumentation* 3.S08007 (2008). Ed. by IOPscience. URL:http://iopscience.iop.org/1748-0221/3/08/S08006

[20] *The MOEDAL Collaboration.* URL:http://home.web.cern.ch/about/experiments/moedal

[21] The CMS Collaboration et al. "The CMS experiment at the CERN LHC". In: *Journal of Instrumentation* 3.08 (2008), S08004. URL: http://stacks.iop.org/1748-0221/3/i=08/a=S08004

[22] CMS Collaboration "The CMS tracker system project: Technical Deign Report", CERN-LHCC-98-006, CMS-TDR-5 - Geneva CERN, (1997).

[23] CMS Collaboration "The CMS electromagnetic calorimeter project: Technical Design Report", CERN-LHCC-97-033, CMS-TDR-4 - Geneva CERN, (1997).

[24] CMS Collaboration, "The CMS hadron calorimeter project: Technical Design Report", CERN-LHCC-97-031, CMS-TDR-2 - Geneva CERN, (1997).

[25] CMS Collaboration, "The CMS muon detector project: Technical Design Report", *Detectors and Experimental Techniques*, CERN, 1997. URL:https://cds.cern.ch/record/343814/fil 97-032.pdf

[26] CMS Collaboration, "The TriDAS Project Technical Design Report, Volume 1: The Trigger Systems", CERN/LHCC 2002/26, CMS Technical Report 6.2, (2002).

[27]V.Gori. "The CMS High Level Trigger", International Journal of Modern Physics: Conference Serie (2014).

[28] CMS Collaboration, "The TriDAS Project Technical Design Report, Volume 2: Data Acquisition and High-Level Trigger", CERN/LHCC 2002/26, CMS Technical Report 6.2, (2002).

[29] *WLCG Project*. URL:http://www.cern.ch/lcg

[30] J. D. Shiers. "The Worldwide LHC Computing Grid (worldwide LCG)", Computer Physics Communications 177 (2007) 219-223.

[31] *European Grid Infrastructure*. URL:http://www.cern.ch/lcg

[32]*Open Science Grid*. URL: http://www.egi.eu/

[33] D. Bonacorsi, "WLCG Service Challenges and Tiered architecture in the LHC era", IFAE, Pavia, April (2006).

[34] *Virtual Organization Membership Service*.

URL: $http://toolkit.globus.org/grid_software/security/voms.php$

[35] *Tier architecture of computing resources*.

URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/WorkBookComputingModel

[36] CMS Collaboration, "The CMS Computing Model", CERN LHCC 2004-035.

[37] CMS Collaboration. "The CMS Computing Project: Technical Design Report", CERN-LHCC-2005-023.

[38] G. Petrucciani, A. Rizzi and C. Vuosalo, on behalf of the CMS Collaboration, "MiniAOD: A New Analysis Data Format for CMS", Journal of Physics: Conference Series 664 (2015) 072052.

[39] *Welcome to Frontier*. URL:http://frontier.cern.ch/

[40] A. Afaq, A. Dolgert, Y. Guo, V. Kuznetsov and al. "The CMS Dataset Bookkeeping Service", Journal of Physics: Conference Series (2008).

[41] M Giffels and Y Guo D Riley. "Data Bookkeeping Service 3 - Providin event metadata in CMS", Journal of Physics: Conference Series (2014).

[42] J. Rehn and T. Barrass and D. Bonacorsi and J. Hernandez and I. Semeniouk and L. Tuura. "PhEDEx high-throughput data transfer management system". In: CHEP06 (2006). Ed. by GridPP.

[43] *CMSSW Application Framework*. URL: https://twiki.cern.ch/twiki/bin/view/CMSPublic/Work

[44] E Fajardo, O Gutsche, S Foulkes et al. “A new era for central processing and production in CMS”,Ed. by IOPscience, In: Journal of Physics: Conference Series.

[45] *CRAB*. URL:https://twiki.cern.ch/twiki/bin/view/CMSPublic/SWGuideCrab

[46] D.Spiga et al. “The CMS Remote Analysis Builder (CRAB)”. FERMILAB-CONF-07-769-CD-CMS (2007).

[47] Tony Wildish and A. Sanchez-Hernandez and R. Egeland and C.-H. Huang and N. Ratnikova and N. Magini. “From toolkit to framework - the past and future evolution of PhEDEx”, Journal of Physics: Conference Series (2012), IOPscience, Vol. 396.

[48] *CMS Dashboard*. URL: http://dashboard.cern.ch/

[49] L. Giommi, “Predicting CMS datasets popularity with Machine Learning”. Thesis, 2015.

[50] Valentin Kuznetsov, Ting Li, Luca Giommi, Daniele Bonacorsi, Tony Wildish. ”Predicting datasets popularity for the CMS experiment”. Cornell University Library, arXiv:1602.07226, 2016.

[51] *DCAF Pilot*.

URL:https://github.com/dmwm/DMWMAnalytics/tree/master/Popularity/DCAFPilot

[52] V. Kuznetsov N. Magini M. Giels Y. Guo and T. Wildish. “ The CMS Data Management System”. In: *J.Phys*(2014)