



CMS Data Quality Monitoring: Offline Workflow and Physics

Sarah Lindner, Karl-Franzens-Universität Graz, Austria

September 8, 2011

Abstract

Data Quality Monitoring plays an important role in the work with CMS data. It ensures that values which are later taken into calculations are reliable. In the following report of my time as a summer student at DESY I will give an insight to the basic concepts of Data Quality Monitoring and explain the projects I worked on as an example for computing and data processing for Data Quality Monitoring.

Contents

1	Introduction	3
2	Data Quality Monitoring (DQM)	3
2.1	From Triggered Signal to Calculation	3
2.1.1	Online DQM	3
2.1.2	Offline DQM	4
2.1.3	Certification and Sign-Off	5
2.2	Helpful Script	5
3	Creation of Histograms	6
3.1	Jet Selection	7
3.2	Restrictions	7
3.3	Resulting Histograms	8
4	Conclusions	12

1 Introduction

During a visit to CERN a few years ago I learned that not every event which occurs in the detectors of LHC is recorded, because it would be way to much data to be stored. In fact most of the data gets thrown away without a single person looking at it, because it doesn't fulfill the requirements set by the "trigger". At that time, I thought: "What are they doing? What if they threw away all the important data for a great discovery?" Now I know that the trigger is set very carefully and does record from time to time "trash" data to make sure there are no loop holes for new discoveries.

When I came to DESY I was told that even more data is sorted out. This is one part of "Data Quality Monitoring (DQM)" work. I worked in the CMS DQM group during my stay at DESY, thus I will write about DQM in section 2.1. In section 2.2 I'm going to describe my first project during this summer school which gives a deeper insight to the work of DQM. Together with my colleague Alexander Chaushev, another summer student, I had to write a program for the so-called "offline shifters" (c.f. section 2.1.2). When this was done, I got a new task by which I learned even more of the work performed by the DQM group. I created histograms for identifying top-antitop pair production. This process and the results I will discuss in section 3.

2 Data Quality Monitoring (DQM)

2.1 From Triggered Signal to Calculation

When a collision takes place there is more data created than can be processed. So as mentioned in the introduction, most of it is already thrown away by an electronic low level trigger. It reduces the event rate of 40MHz to 200Hz (0.25%) [1].

After this the process of "Data Quality Monitoring (DQM)" begins, which is a part of data processing as described in the subsequent paragraphs [2]. DQM shifting, which ensures the accuracy of data is split into online shifts which take place 24/7 at the CMS site at CERN and offline shifts at CERN, DESY and FNAL. Usually, one offline shifter works for 1-2 weeks, 6 hours a day. At DESY the shifts take place in the CMS center in building 1a. Afterwards, certification and sign-off is performed for final validation.

2.1.1 Online DQM

After having passed the low-level trigger filter the next step is fulfilling the requirements of the high level trigger filter. At this time the event rate is reduced to 10-15 Hz. From there data in the form of histograms is sent to a storage manager proxy server, which saves the histograms into different files, according to what they represent. Beginning at this point, certain applications can be run on the data. For example, there exist algorithms checking hot, cold or otherwise bad channels, noise levels, occupancy, timing problems, trigger issues, detector-specific known problems etc. The data then is put on the so-called DQM GUI (graphical user interface, c.f. figure 1) where it is visualized including alarm states. Data which got to this stage is then stored on disc and when a

certain amount of data is reached, it is backed up on tape. For later re-inspection DQM data is always kept on disc for several months.

2.1.2 Offline DQM

In the CMS offline DQM system histograms are filled and stored as "run products" at first. Then, in a procedure called "harvesting" all histograms of one type of one run are summed together. Thus full statistics on the dataset is obtained. What the harvesting algorithm also does is performing the preliminary automatic data certification decision. All information gathered so far (histograms, certification results, quality test results) are stored as ROOT files [3] and put on the DQM GUI web server. As in online DQM, data is stored on disc after having reached specified size archived on tape. The results from the automatic harvesting certification are uploaded on the "run registry", see figure 2. The run registry is a database where on the one hand the workflow is controlled by a user interface which tells what data is available. On the other hand, information is stored there.

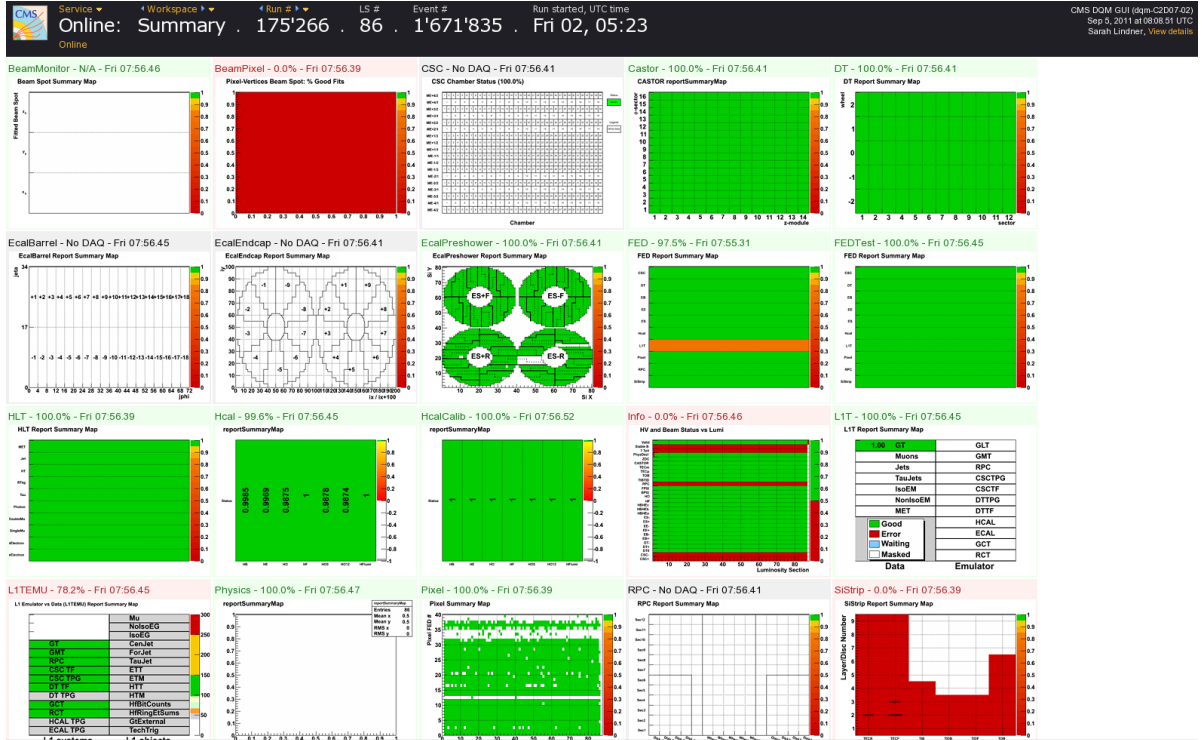



Figure 1: Extract of the DQM GUI showing tracker information. Green means the inspected part of the detector part works as expected, red indicates a defect



CMS DQM Run Registry (Global)


GLOBAL

Analysis

Tools

Login

Runs



RunInfo

Refresh

Table

LumiSec

17,334 items. Show 20 from 301 to 320. Page 16 / 86

Run Number	Group	Events	Rate, Hz	Run Started	Run Duration	LS	E	Fill	L1(124)	SI Bk	State	Dataset	Shifter	CASTOR	CSC	DT	ECAL	ES	HCAL	HLT	L1T	
172799	Collisions11	408854886	47031.392	Fri 05:08:11 18:11:00	00:02:25:00	372	3500.04	2006	4216872	✓	SIGNOFF	/Global /Online/ALL	Marko Kovac	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	
											COMPLETED	/PromptReco /Run2011A-PromptDQM	Terence Libeiro	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	
172798	Collisions11	31948272	38272.579	Fri 05:08:11 17:36:00	00:00:14:00	35	3500.04	2006	173396	✓	SIGNOFF	/Global /Online/ALL	Marko Kovac	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	
											COMPLETED	/PromptReco /Run2011A-PromptDQM	Hannelies Kluge	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	
172791	Collisions11	2211799637	57217.586	Fri 05:08:11 07:00:00	00:10:47:00	1664	3500.04	2006	1108499	✓	SIGNOFF	/Global /Online/ALL	Ulysses Grunder	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	
											COMPLETED	/PromptReco /Run2011A-PromptDQM	Hannelies Kluge	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	
172789	Commissioning11	1770784	1163.842	Fri 05:08:11 06:31:00	00:00:24:55	63	3500.04	2006		1	✗	SIGNOFF	/Global /Online/ALL	Franciele Da Cunha Marinho	GOOD	BAD!	BAD!	GOOD	BAD!	GOOD	GOOD	GOOD
172784	Commissioning11	406361	575.209	Fri 05:08:11 05:54:00	00:00:11:37	29	3500.04	2006		0	✗	SIGNOFF	/Global /Online/ALL	Franciele Da Cunha Marinho	GOOD	BAD!	BAD!	GOOD	BAD!	GOOD	GOOD	GOOD
172782	Commissioning11	1020809	768.377	Fri 05:08:11 05:23:00	00:00:22:26	56	3500.04	2006		0	✗	SIGNOFF	/Global /Online/ALL	Franciele Da Cunha Marinho	GOOD	GOOD	BAD!	GOOD	BAD!	GOOD	GOOD	GOOD
172780	Cosmics11	5241238	887.629	Fri 05:08:11 03:31:00	00:01:39:00	254	3500.04	2006		0	✗	SIGNOFF	/StreamExpress /Cosmics11-Express/DQM	Amr Radi	GOOD	GOOD	GOOD	GOOD	BAD!	GOOD	BAD!	GOOD
												SIGNOFF	/Global /Online/ALL	Franciele Da Cunha Marinho	GOOD	GOOD	GOOD	GOOD	BAD!	GOOD	GOOD	GOOD
172778	Collisions11	112109755	48249.047	Fri 05:08:11 02:25:00	00:00:38:00	98	3500.04	2005	1021	✓	SIGNOFF	/Global /Online/ALL	Franciele Da Cunha Marinho	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	
											COMPLETED	/PromptReco /Run2011A-PromptDQM	Amr Radi	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	GOOD	

Figure 2: Extract of the run registry for several runs with flags for each detector part

2.1.3 Certification and Sign-Off

When the automatic certification results are uploaded on the run registry, detector and physics quality is evaluated by people who are on shift. Thus the automatic results are checked once more by a human. Ideally last remaining errors are wiped out in this process, i.e. only good data is kept.

The final step is confirmation by the detector and physics object groups. For this purpose, regular sign-off meetings take place.

2.2 Helpful Script

Alexander and I were asked to write a Python script which checks if there are any runs that have to be certified by the offline shifter. The idea was that this program runs repeatedly after a certain time has passed and send an automatized message if there is a new run to be looked at. Thus the shifters don't have to remember themselves to check for new runs every now and then and go through the whole process of finding a run which has to be certified by them.

The program can be divided into two parts, one which checks if runs were already certified by the online shifters, the other one checking whether these runs were already put on the run registry by offline shifters.

The GUI part of our program which makes sure if the runs certified by the online shifters does this by checking if they are already uploaded to the DQM GUI. For accessing the documents on the GUI we used the bash command “curl”, which we implemented in our Python program. As a double check the program not only queries the uploaded runs, but also tests whether all important files got uploaded along with it.

For the accessing the runregistry there is an example script on the CMS “TWiki”, the CMS documentation platform. It also shows how to set the filter options [4]. With the filter options it is possible, to query for specific parameters, like date, run number, group name, name of the dataset. We selected runs created by collisions (on the contrary to events caused by cosmic rays).

In the end, the results of these functions are compared and the runs fulfilling all the requirements get output into a file together with the date and time when the program found it. The interval for checking for new runs can be changed in the script.

One problem we faced was that every time the program accessed the GUI or the run registry it asked for the user’s password for their CERN certificate. This contradicted our idea of a program which runs by itself. Finally we solved the problem by setting up a proxy certificate.

3 Creation of Histograms

As can be guessed when looking at figure 1, there is a broad variety of histograms on the DQM GUI. Which types of histograms appear on the GUI is constantly updated. For this purpose, new ones have to be created that can later be filled with data from the latest events. With the guidance of my supervisor Andreas Meyer I edited a ROOT [3] program which puts out histograms of a top-antitop pair decay. The event I was looking for is a single muon, four jet event as sketched in figure 3.

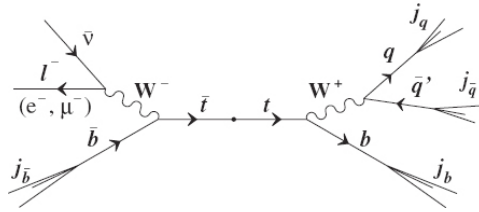


Figure 3: Sketch of a top-antitop pair decaying into a single lepton and four jets

In the following paragraphs I will explain and illustrate the steps necessary to extract events which are candidates for a top-antitop decay. The basic idea is to look for events where one muon was created and add up the masses of the three jets produced by one

of the two top particles. If the outcoming is about 175 GeV, it is assumed that an event in which a top and an antitop particle were produced was found.

3.1 Jet Selection

As mentioned, the mass of the three jets caused by one top are added up. Unfortunately, there are at least four jets created in the considered events. The way out is to first make the vectorial sum of the momentum of all possible combinations of three jets. We assume that the jets created by the same particle travel in a small angle relative to each other. Thus if the vectorial sum of the momentum is highest, it is most probable that these jets originate from the same particle. After having found the desired jets, the mass of the combination of them can be calculated. But to get a reasonable result, there are some restrictions to be made.

3.2 Restrictions

To get an obvious peak and to reduce the background, some following constraints were imposed. With the exception of the cuts they were already implemented in the original program which I edited, so I will just shortly illustrate what some of them are doing.

- **Jet energy scale correction:** When particles decay into quarks, the quarks immediately hadronize, forming jets. The detected energy of the isn't the real energy of the quark, because of several reasons. Some examples are: the energy of the hadrons is affected by interactions with detector material, some of the particles of the jets get lost on the way to the detector, neutrinos can be formed, which can't be detected. The jet energy scale correction's aim is to get rid of these errors. It is a crucial function for getting a right distribution, that's why included it in all of the shown histograms.
- **Cuts:** On the muon transverse momentum (pt) and the jet transverse momentum cut were applied. This means that only those muons were taken into account, whose momentum was bigger than 25 GeV and jets with a momentum bigger than 30 GeV. The muon cut is needed to suppress muons which are formed by another process than the W decay. These other muons usually have lower transverse momentum. With the cut on the jets, higher order jets are concealed which don't originate from either the bottom quark or the quarks created by the W decay, but from a particle which has been created in the primary quark jet.
- **Muon Isolation:** Checks that the muon doesn't travel in the same direction as other particles, which would imply that the muon doesn't originate from the W boson.
- **Electron Selection:** Gets rid of electrons which were identified as jets.

- **Jet-electron overlap:** If the angle between an analyzed jet and an analyzed electron is very small, they cannot be treated separately. Thus these events get suppressed.

3.3 Resulting Histograms

The program for creating the histograms was run with data of about 8000 events produced by a Monte Carlo simulation. Unfortunately, due to technical problems, there wasn't more data available for me at the time I made the histograms. With more data the histograms would have been more accurate.

When looking at the following pictures the effects of the applied restrictions can be seen. The first three graphs (figures 4-6) don't include any of the restrictions mentioned in section 3.2 except for the jet energy scale correction. In the histogram of the transverse momentum of the three selected jets a broad peak arises at a little bit less than 150 GeV. In the histogram of the mass a peak is situated at about 150 GeV. This value is too low compared to the 175 GeV we are looking for.

When the cuts are applied (figures 8-9), the number of events fulfilling the requirements drops to less than 20%. The peaks also moved to a position of higher energy. This indicates that muons and jets created by other, secondary processes with lower energy were gotten rid of.

We expected the peak for the mass when all restrictions are set to be at about 175 GeV, which would be the mass of the top quark, but it appears at slightly more than that, at about 180 GeV. The reason for this can be that the data simulation was performed with a value greater than 175 GeV.

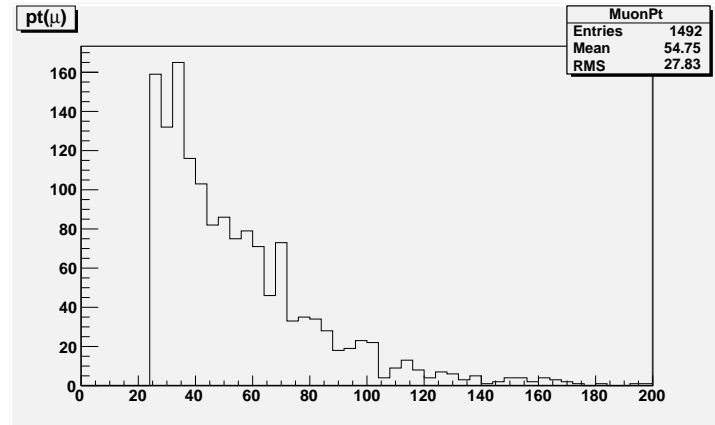


Figure 4: Transverse momentum of the muon without restrictions

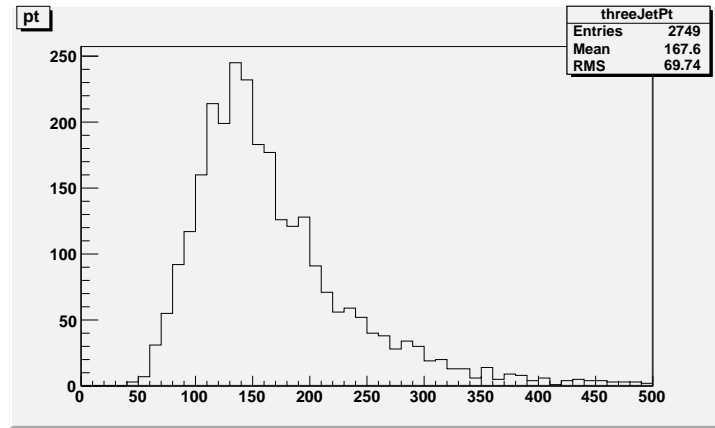


Figure 5: Transverse momentum of the three jets without restrictions

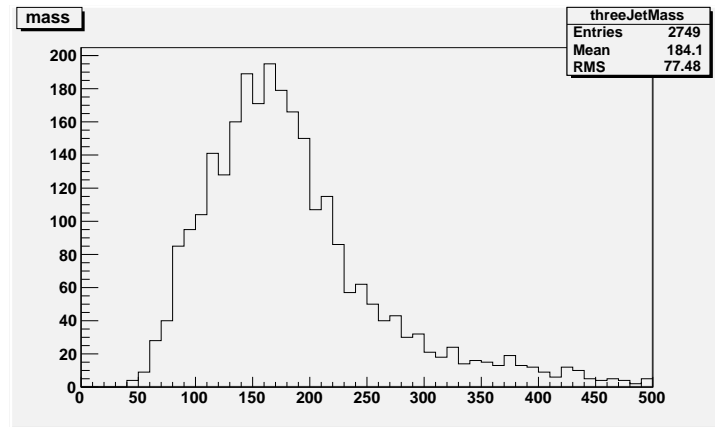


Figure 6: Mass of the three jets without restrictions

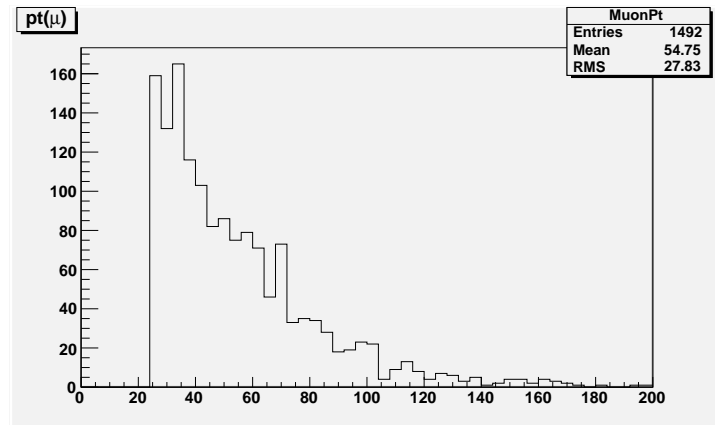


Figure 7: Transverse momentum of the muon with muon $pt > 25$ GeV and jet $pt > 30$ GeV

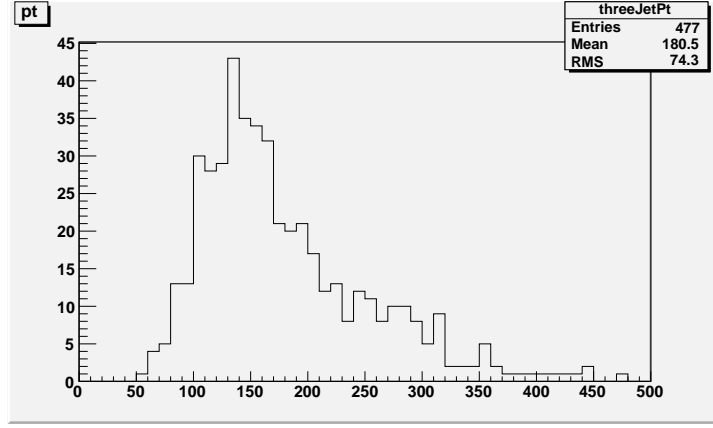


Figure 8: Transverse momentum of the three jets with muon $pt > 25$ GeV and jet $pt > 30$ GeV

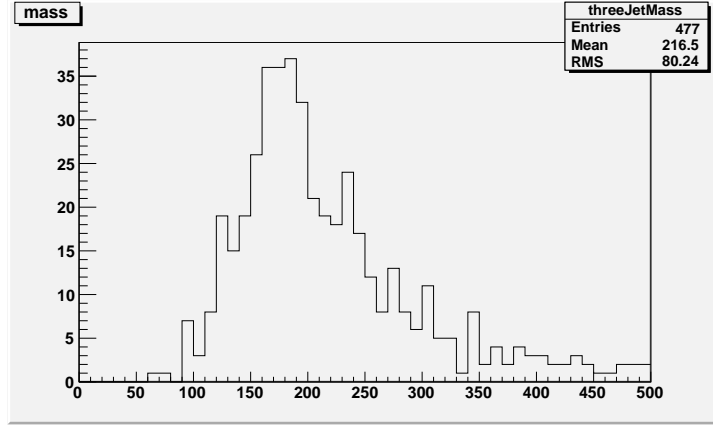


Figure 9: Mass of the three jets with muon $pt > 25$ GeV and jet $pt > 30$ GeV

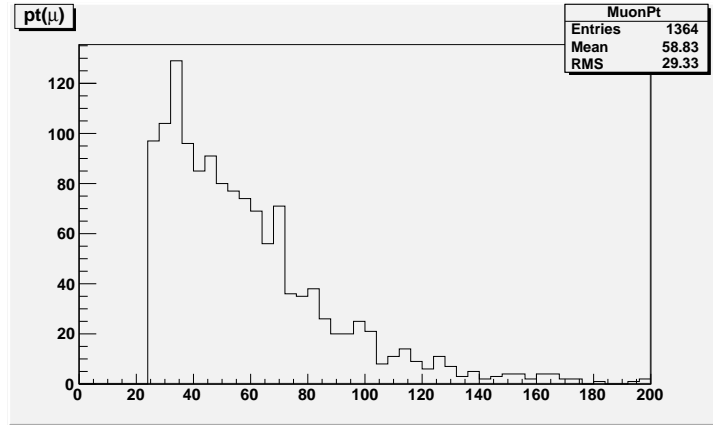


Figure 10: Transverse momentum of the muon with all restrictions

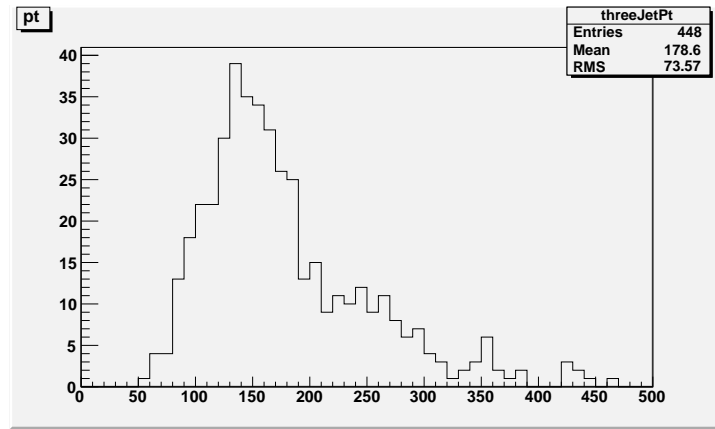


Figure 11: Transverse momentum of the three jets with all restrictions

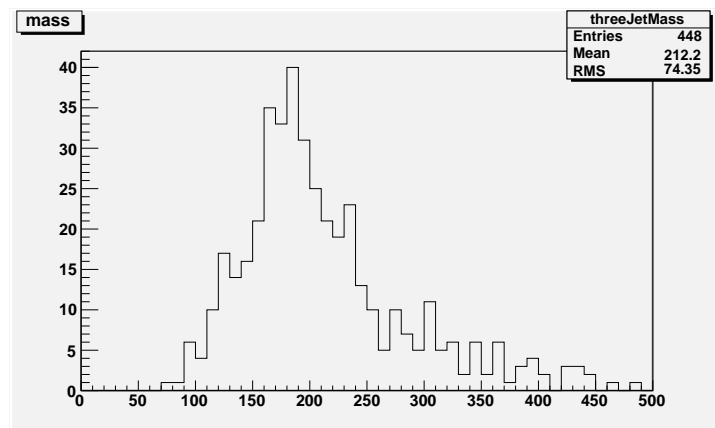


Figure 12: Mass of the three jets with all restrictions

4 Conclusions

By working on the two described projects, I got an insight to the field of work of Data Quality Monitoring in CMS. One important task is making sure that calculations are not carried out on data which was taken by defect detectors or otherwise bad data. Another responsibility is to merge raw data to histograms showing physical properties.

References

- [1] LHC-Computing: Vom Detektorsignal zur Ereignisinterpretation *M. Edelhoff, T. Kreß*
- [2] CMS Data Quality Monitoring: systems and experiences *L. Tuura, A. Meyer, I. Segoni, G. Della Ricca*
- [3] ROOT - A data analysis framework, 2009, <http://root.cern.ch>
- [4] CMS DQM Run Registry API,
<https://twiki.cern.ch/twiki/bin/viewauth/CMS/DqmRrApi>
- [5] Central DQM Shift Tutorial Online/Offline,
<https://indico.cern.ch/getFile.py/access?resId=1&materialId=slides&confId=148522>

List of Figures

1	DQM GUI	4
2	DQM run registry	5
3	Top-antitop decay; J. A. R. Cembranos, A. Rajaraman and F. Takayama: Searching for CPT violation in $t\bar{t}$ production	6
4	Transverse momentum of the muon without restrictions	8
5	Transverse momentum of the three jets without restrictions	9
6	Mass of the three jets without restrictions	9
7	Transverse momentum of the muon with muon $pt > 25$ GeV and jet $pt > 30$ GeV	9
8	Transverse momentum of the three jets with muon $pt > 25$ GeV and jet $pt > 30$ GeV	10
9	Mass of the three jets with muon $pt > 25$ GeV and jet $pt > 30$ GeV	10
10	Transverse momentum of the muon with all restrictions	10
11	Transverse momentum of the three jets with all restrictions	11
12	Mass of the three jets with all restrictions	11