Lecture 9: Data Mining, Data Analytics and Big Data

Maaike Limper, Antonio Romero, Manuel Martin

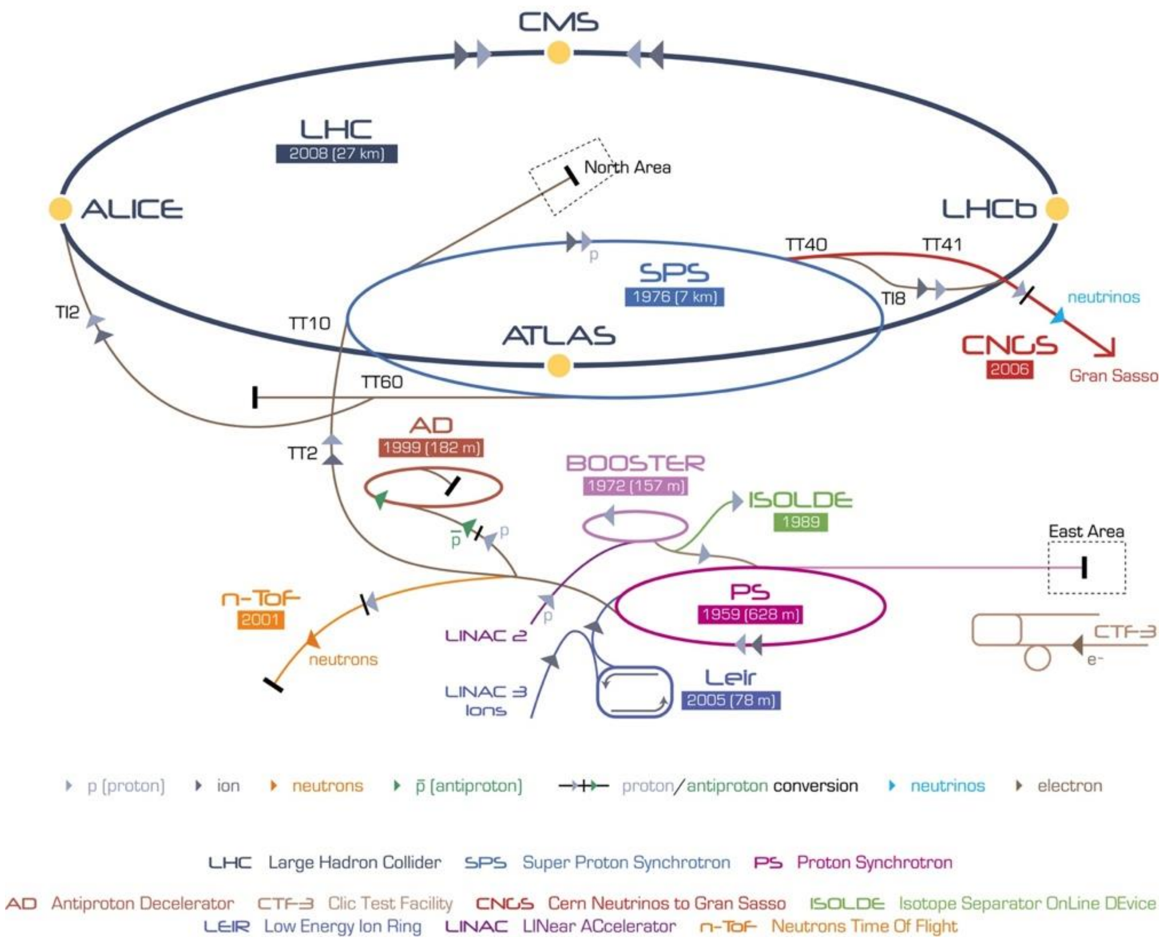
Introduction

- Two openlab Projects in IT-DB
  - Data Analytics
  - In-Database Physics Analysis
- Both using data analytics
- Projects have different scopes

Summary

- CERN environment
- Data Analytic Project
- R and Oracle R
- Data Discovery

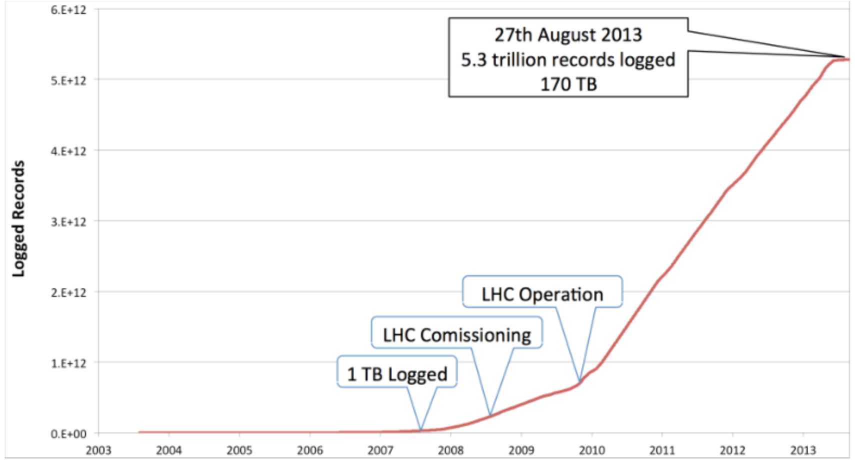
CERN is a extreme data environment



- Control and operations
  - Million of sensors, signals
  - Large number of control devices
  - Equipment
- Monitoring and logging
- Supporting IT infrastructure
  - Databases
  - Network
  - Services

CERN Data Investment

- CERN generates huge amount of data everyday
  - Accelerator Logging Service around 275 GB/day
- CERN has great monitoring and logging systems
  - Large amount of data has been stored over years



eLogbook AB & OP

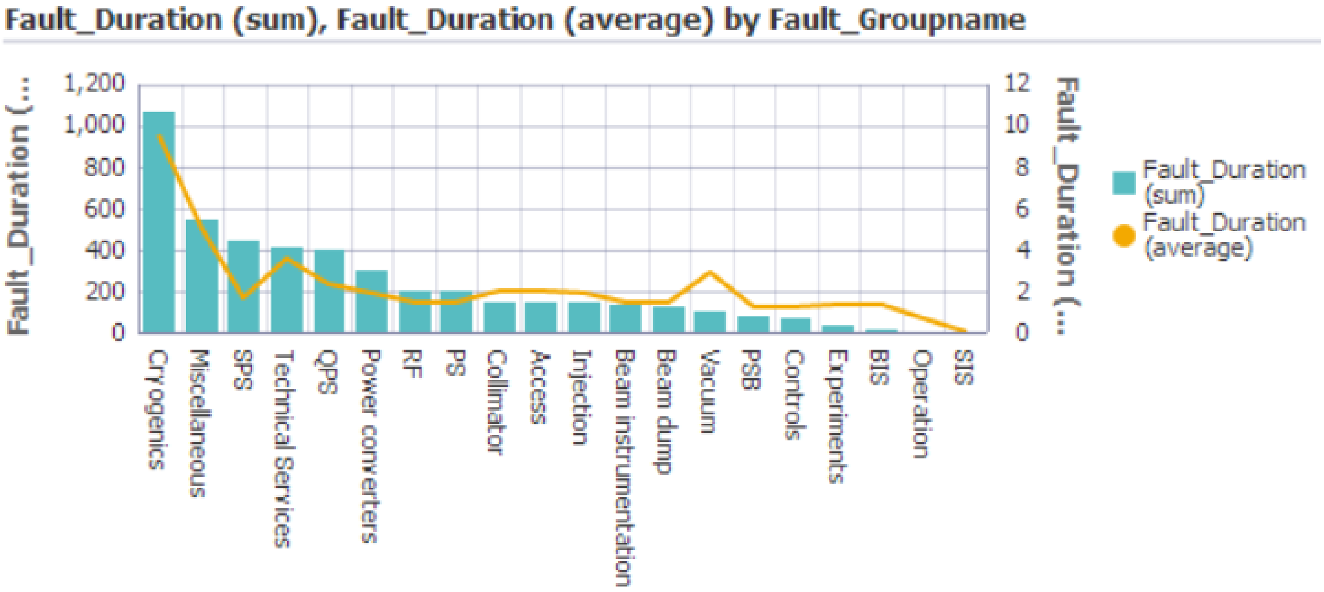
timber The LHC Logging System

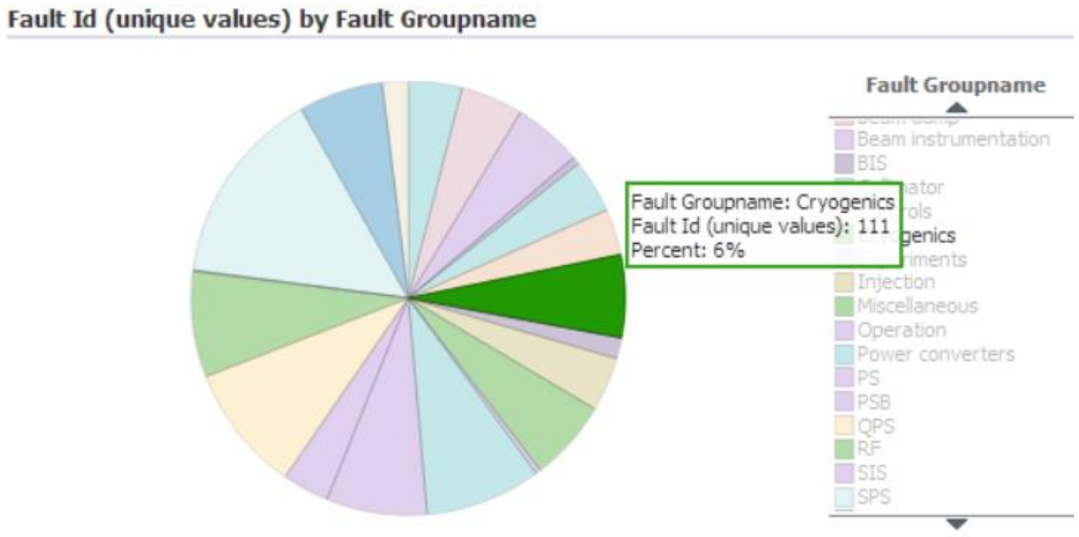
CASTOR CERN Advanced STORage manager

DIAMON

Faults

- Some faults cannot be avoid
- Decrease the availability for running physics

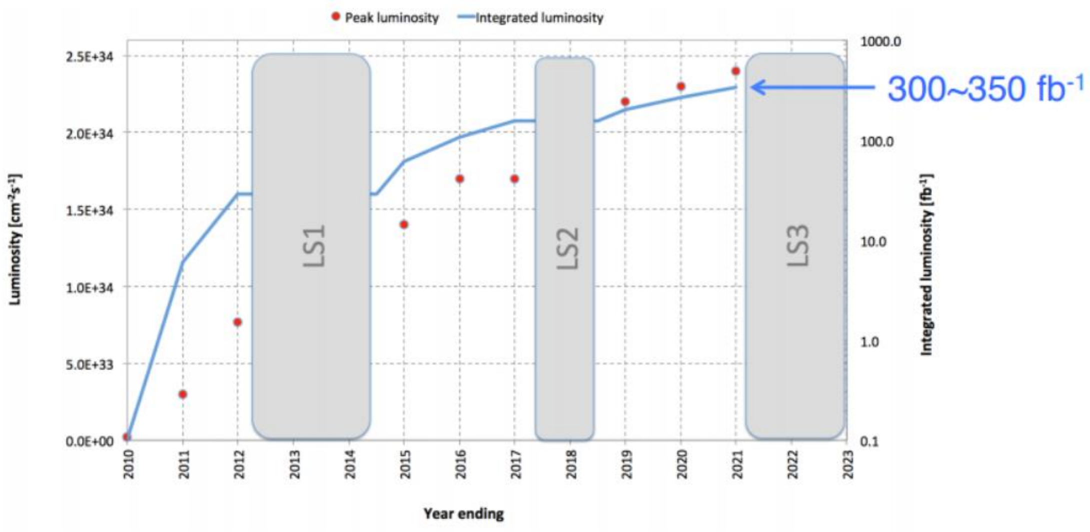




A look into the Future

- Future LHC upgrades will increase luminosity
  - Computing resources needs will be higher
  - Data generated will increase drastically

Parameter	2010	2011	2012	design value
Beam energy	3.5	3.5	4	7
β* in IP 1 and 5 (m)	2.0/3.5	1.5/1.0	0.6	0.55
Bunch spacing (ns)	150	75/50	50	25
Max. number of bunches	368	1380	1380	2808
Max. bunch intensity (protons per bunch)	1.2×10^{11}	1.45×10^{11}	1.7×10^{11}	1.15×10^{11}
Normalized emittance at start of fill (mm mrad)	≈2.0	≈2.4	≈2.5	3.75
Peak luminosity ($cm^{-2}s^{-1}$)	2.1×10^{32}	3.7×10^{33}	7.7×10^{33}	1×10^{34}
Max. mean number of events per bunch crossing	4	17	37	19
Stored beam energy (MJ)	≈28	≈110	≈140	362



- Next accelerators
  - Future Circular Collider (80-100 km)

The objective – Improve our systems

Monitoring and Diagnostics Systems

Data Analytics

Predictive and Proactive systems

Summary

- CERN environment
- Data Analytic Project
- R and Oracle R
- Data Discovery

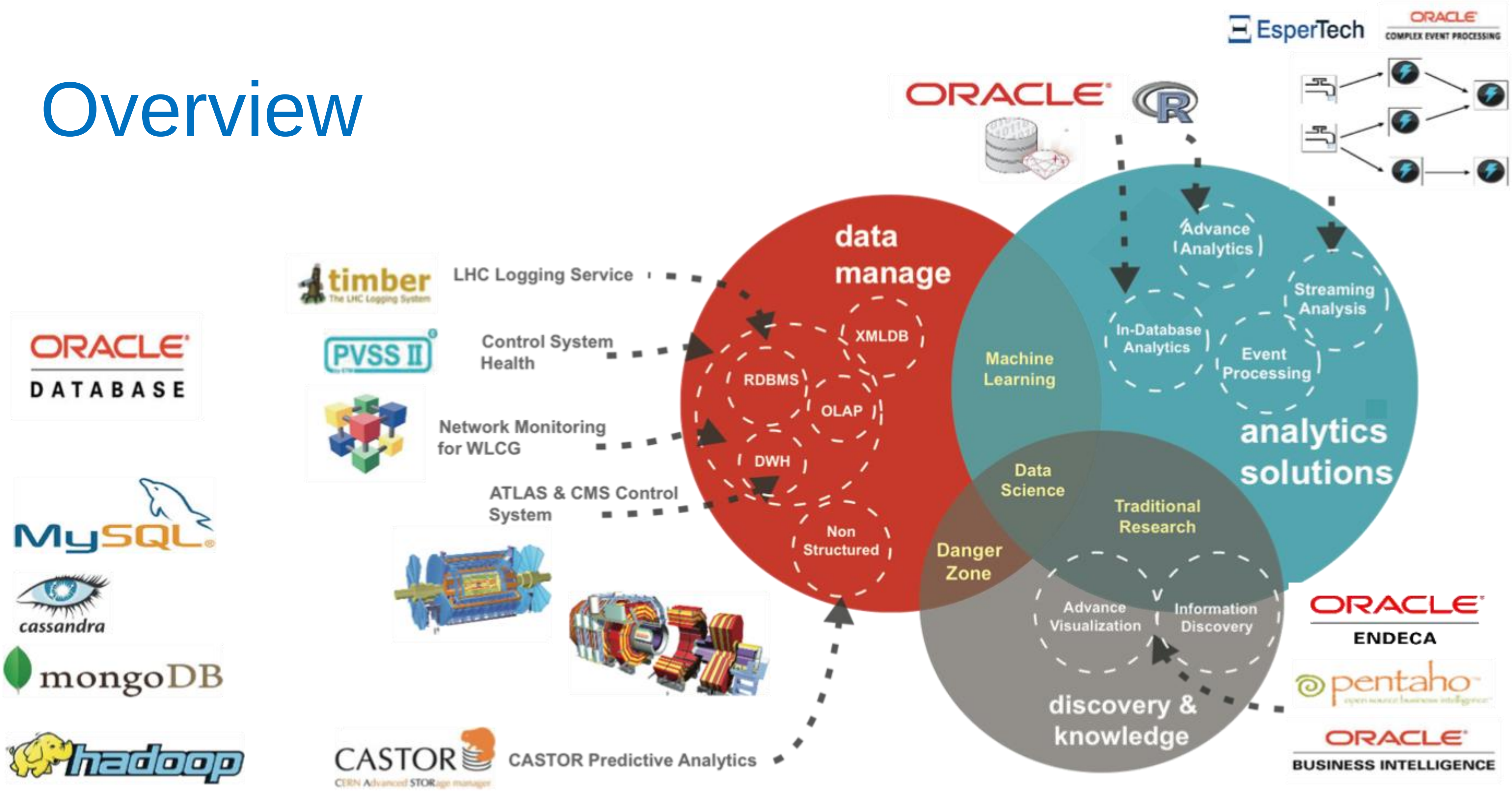
openlab Data Analytics Project

- Optimize our systems
  - Reducing and predicting faults and corrective interventions
  - Increase the availability and operations efficiency
- Profit from CERN data investment by using data analytics
  - Extract knowledge
  - Discover useful information
  - Suggest conclusions
  - Support decision making
- Control and Monitoring Systems
  - Proactive
  - Predictive
  - Intelligent

Data analytic challenges at CERN

- Very dynamic and heterogeneous environment
  - Projects (number of projects)
  - Requirements (data analytic needs)
  - Technologies (problem-driven)
  - Data sources (raw, structured and unstructured)
- Large amount of data
- Education and Training
  - Users know the data and the questions
  - Help them how to connect both

Overview



ORACLE DATABASE
MySQL
cassandra
mongoDB
hadoop
timber
The LHC Logging System
LHC Logging Service
PVSS II
Control System Health
Network Monitoring for WLCG
ATLAS & CMS Control System
CASTOR
CASTOR Predictive Analytics
data manage
XMLDB
RDBMS
OLAP
DWH
Non Structured
Machine Learning
Data Science
Danger Zone
Traditional Research
Advance Analytics
Streaming Analysis
In-Database Analytics
Event Processing
analytics solutions
Advance Visualization
Information Discovery
discovery & knowledge
ORACLE
EsperTech
ORACLE COMPLEX EVENT PROCESSING
ORACLE ENDECA
pentaho
ORACLE BUSINESS INTELLIGENCE

Summary

- CERN environment
- Data Analytic Project
- R and Oracle R
- Data Discovery

What is R

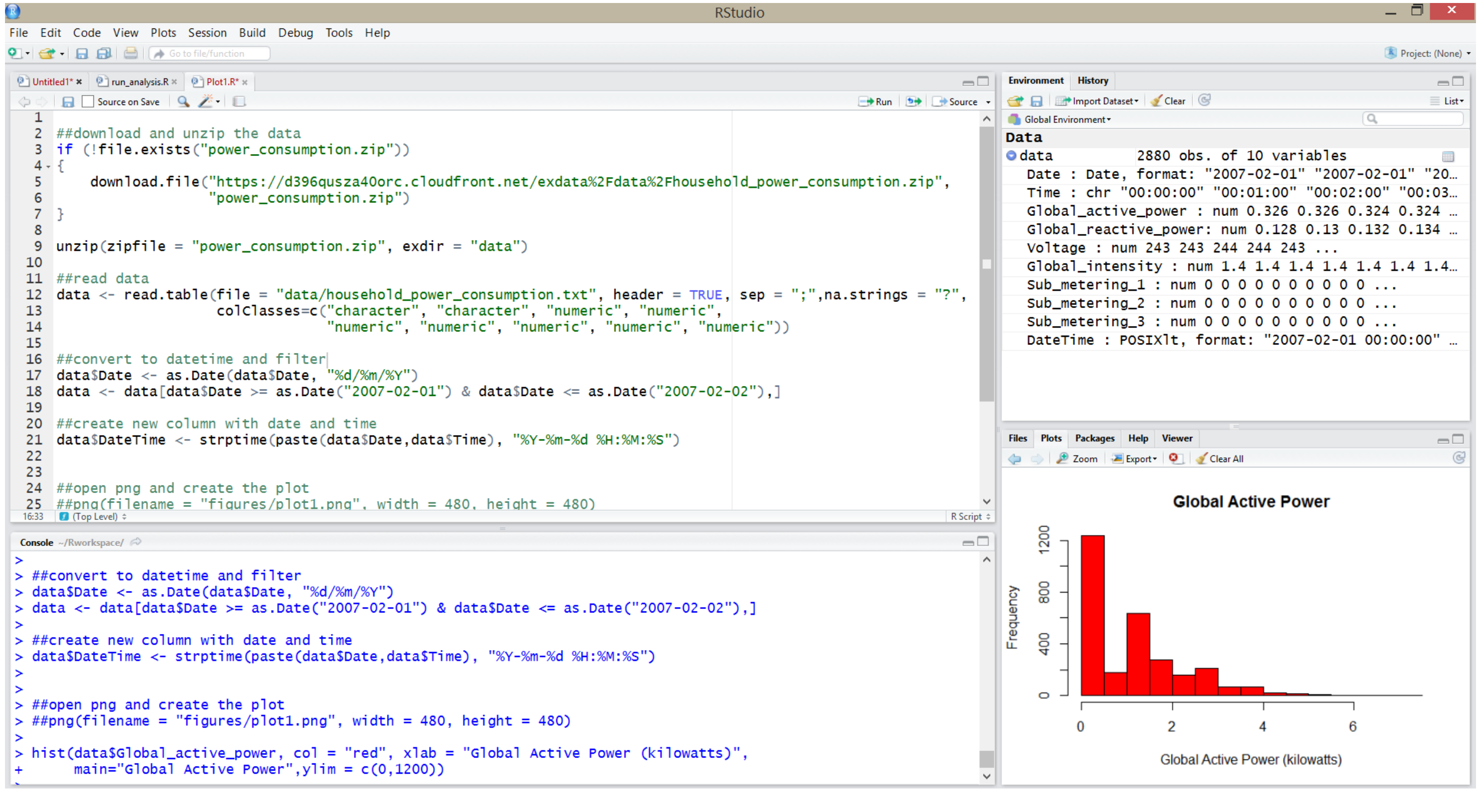
- Language for statistical computing and graphics
- Standard and advanced statistical techniques
- Integrated suite of software facilities
  - Data manipulation
  - Calculation
  - Graphical display
- Free and Open Source

Why R is good for data analysis

- Powerful
  - Advanced statistics
  - Plotting
- Extensible
  - Over 5800 CRAN packages extending base functionality
- Standard de facto for data analytics
- Great user community
  - Over 2 million R users worldwide
- Active development, frequently updated

IDE for R

- Multiple IDEs for R
- RStudio
  - Open source version
  - Windows, Mac, Linux
  - Web (RStudio server)



R example

```
> head(airquality)
  Ozone Solar.R Wind Temp Month Day
1    41     190  7.4   67     5   1
2    36     118  8.0   72     5   2
3    12     149 12.6   74     5   3
4    18     313 11.5   62     5   4
5    NA      NA 14.3   56     5   5
6    28      NA 14.9   66     5   6
```

New York Air Quality Measurements

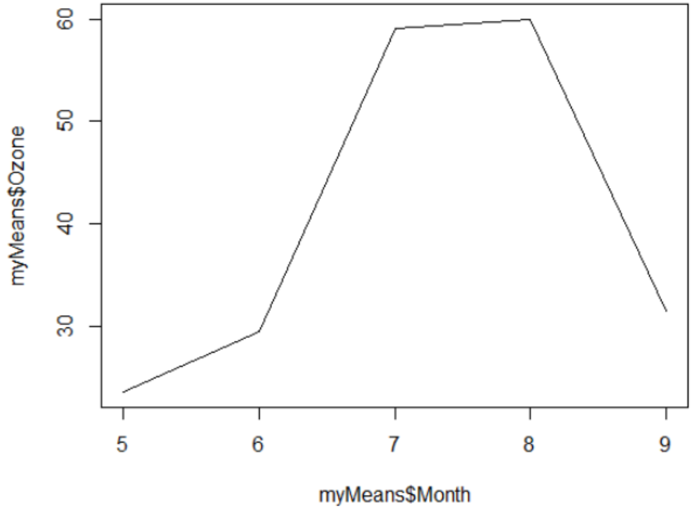
```
> summary(airquality)
     Ozone           Solar.R           Wind             Temp           Month            Day
 Min.   :  1.00   Min.   :  7.0   Min.   : 1.700   Min.   :56.00   Min.   :5.000   Min.   : 1.0
 1st Qu.: 18.00   1st Qu.:115.8   1st Qu.: 7.400   1st Qu.:72.00   1st Qu.:6.000   1st Qu.: 8.0
 Median : 31.50   Median :205.0   Median : 9.700   Median :79.00   Median :7.000   Median :16.0
 Mean   : 42.13   Mean   :185.9   Mean   : 9.958   Mean   :77.88   Mean   :6.993   Mean   :15.8
 3rd Qu.: 63.25   3rd Qu.:258.8   3rd Qu.:11.500   3rd Qu.:85.00   3rd Qu.:8.000   3rd Qu.:23.0
 Max.   :168.00   Max.   :334.0   Max.   :20.700   Max.   :97.00   Max.   :9.000   Max.   :31.0
 NA's   :37       NA's   :7
```

R example

```r
library(plyr)

## subset needed data
myData <- airquality[,c("Month","Ozone","Solar.R","Wind")]

##group by Month and calculate means
myMeans <- ddply(.data = myData,
                 .variables= .(Month),
                 .fun= numcolwise(mean, na.rm = TRUE))

## show calculated dataset
myMeans
```

```r

plot(x = myMeans$Month, y = myMeans$Ozone, type = "l")

```

```
myMeans
Month      Ozone  Solar.R      Wind
    5 23.61538 181.2963 11.622581
    6 29.44444 190.1667 10.266667
    7 59.11538 216.4839  8.941935
    8 59.96154 171.8571  8.793548
    9 31.44828 167.4333 10.180000
```



Some R Resources

- CRAN (http://cran.r-project.org)
- R Tutorial (http://www.r-tutor.com/)
- R-bloggers (http://www.r-bloggers.com/)

- Free courses
  - Coursera - “Data Science” (Johns Hopkins University)
  - Lynda - “Up and Running with R”
  - O’Really - “Try R” (http://tryr.codeschool.com)

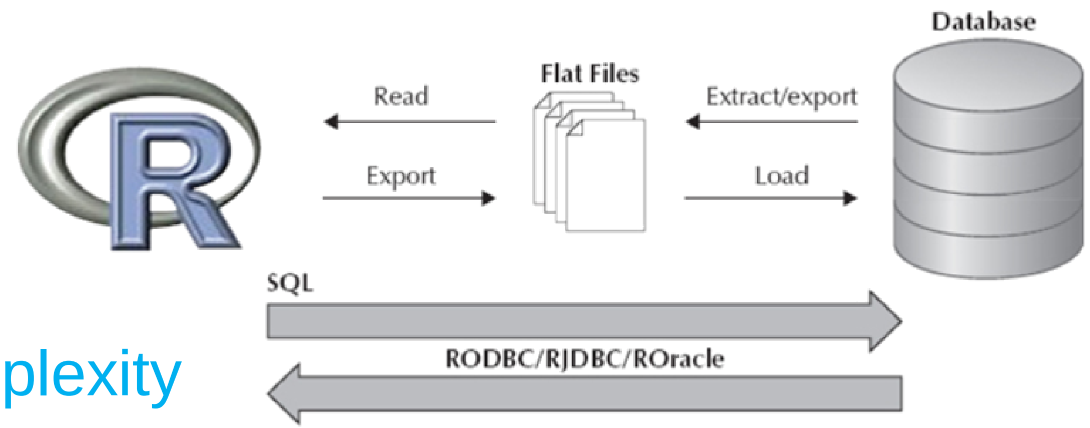
- Learn R, in R
  - Swirl (http://swirlstats.com)

{swirl}

Learn R, in R.

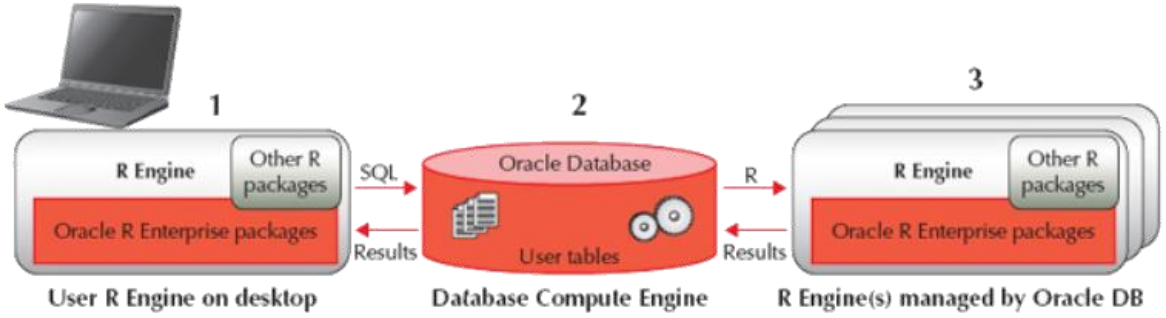
Traditional R and Database Interaction

- Data has to be moved from database to client
  - Client need to store it in HDD
- R user needs to know SQL
  - Mixing R and SQL
- Not multithreaded or parallel
  - Use special packages increasing complexity
- R client has to load everything in memory



Oracle R Enterprise

- A database-centric environment for analytical processes in R
  - Allows to use the database server to run R scripts
  - R working on data directly in the database
  - Integration with the SQL language
    - Run R scripts from SQL



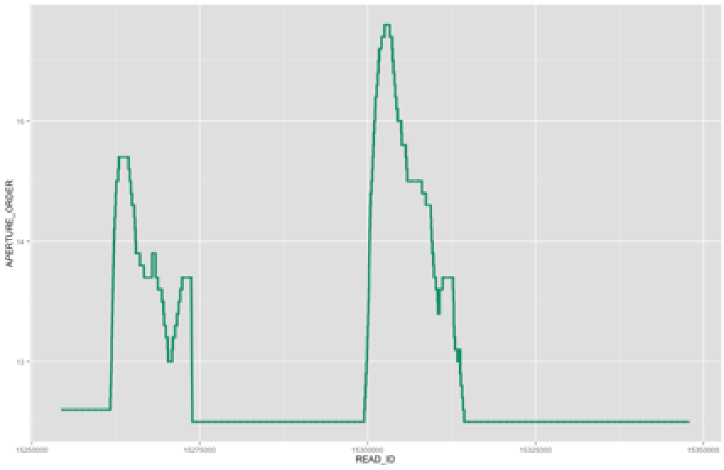
- Transparency Layer
  - Transparently interact with the database in R

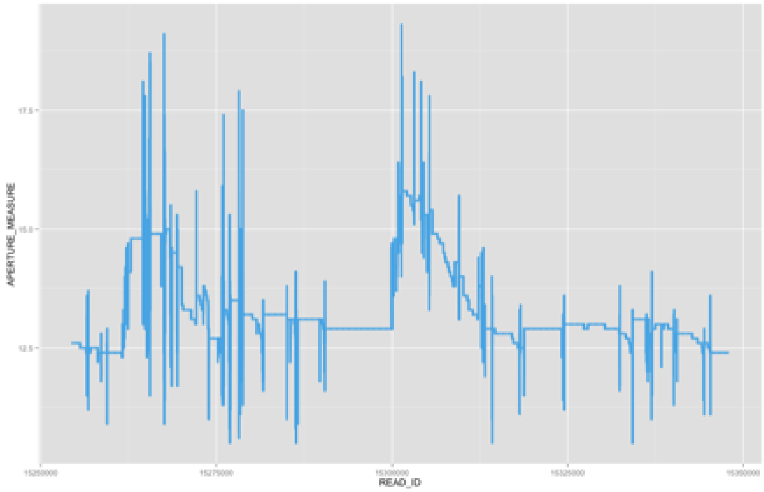
Benefits of using ORE

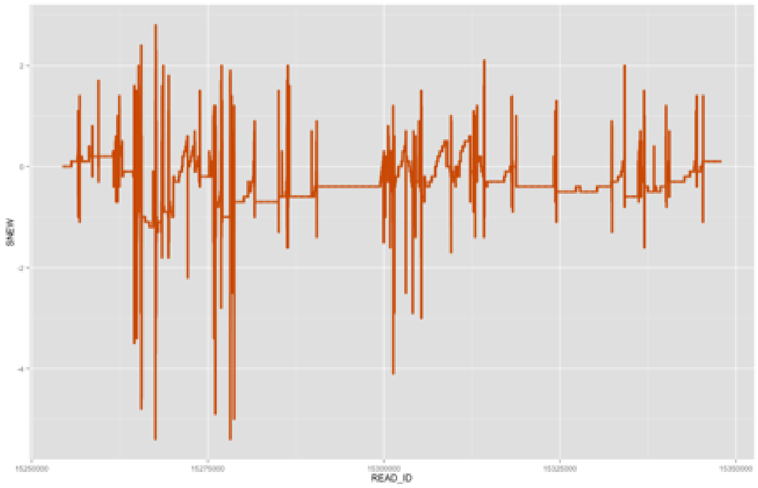
- Allows in-database data analysis
- Provide data parallelism and resource management
- Execute R scripts in database server machine
  - Scalability and performance
  - Eliminate memory constraint
- Oracle databases are widely used at CERN

Cryogenics Use Case: Faulty Valves Detection

- What is the objective?
  - Predict faulty valves before they actually fail
- How?
  - Valve receive an aperture order value (**aperture order**)
  - Effective aperture realized by the valve (**aperture measured**)
  - Analyzing the difference between both (**S** = **aperture order - aperture measured**)

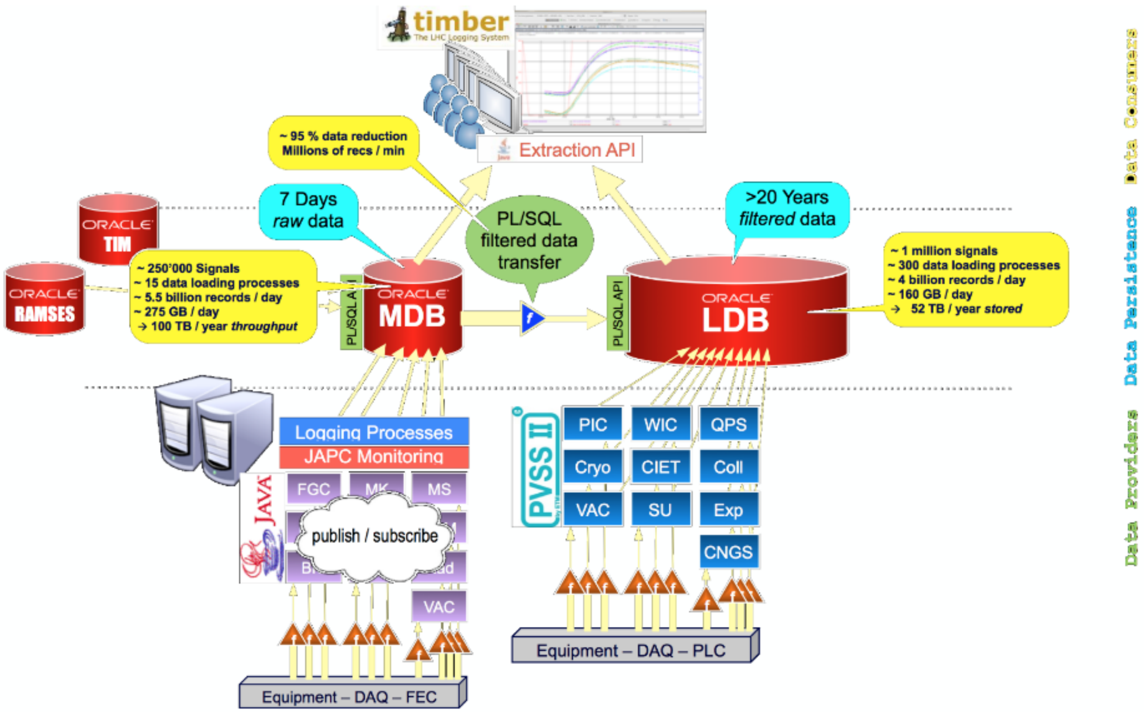






Cryogenics Use Case: Faulty Valves Detection

- Signals used
  - S = aperture order - aperture measured
- Features extractions based on S
  - Variance
  - Percentile 99.9
  - Rope distance
  - Noise Band
- **Automatic Faulty Valves Detection System**
  - SVM - Support Vector Machine
- The learning set 44 valves
- Three different status
  - Faulty,
  - Not faulty
  - Unknown
- Data comes from Accelerator Logging Service



Summary

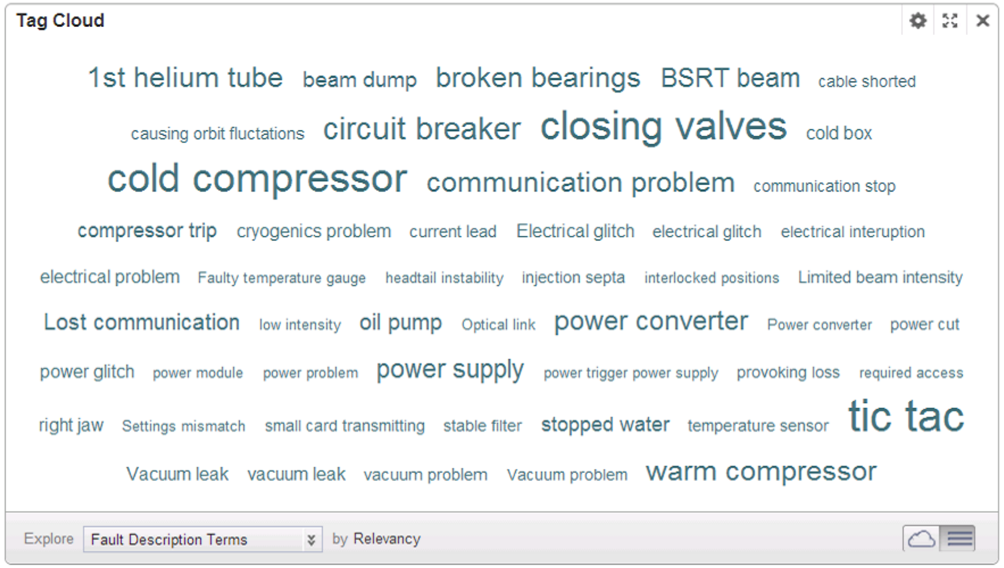
- CERN environment
- Data Analytic Project
- R and Oracle R
- Data Discovery

Data Discovery

- Interactive and visual analytics
- Intended to be used by the end users
  - Enabling them to use their intuition and knowledge of the data
- Powerful customization of dashboards and visualizations
  - Without intervention of IT
- Structure and unstructured data
- Mainstream field in Data Analytics

Endeca Information Discovery

- Data discovery platform
  - Analyze information of any type and any source
- Flexible and user friendly
- Professional ETL tool
- Powerful text analysis
  - Sentiment
  - Text tagging, entities extraction
  - Multi language features



Endeca Use Case

- Data Discovery for **improving** the Accelerator Complex Operations
- Electronic Logbook
  - Log of events in the accelerator complex

Statistics for the eLogbook: PS From: 20120815 Period: Morning To: 20120817 Period: Morning

Start Statistics ...

Availabilities			
Lines	In Super Cycle	In Fault	Availabilities (%)
AD	55 [h]	7 [h] 57 [min] 37[s]	85%
EASTA	55 [h]	7 [h] 57 [min] 37[s]	85%
EASTB	55 [h]	7 [h] 57 [min] 37[s]	85%
SFTPRO	6 [h] 54 [min] 29[s]	5 [h] 45 [min] 18[s]	16%
EASTC	55 [h]	7 [h] 57 [min] 37[s]	85%
CNGS	55 [h]	10 [h] 22 [min] 58[s]	81%
LHCPROBE	55 [h]	7 [h] 57 [min] 37[s]	85%
I_LHC	31 [h] 31 [min] 33[s]	4 [h] 37 [min] 40[s]	85%
LHC	55 [h]	7 [h] 57 [min] 37[s]	85%
TOTAL	423 [h] 26 [min] 02[s]	68 [h] 31 [min] 38[s]	83%

Systems for AD		
GROUP NAME	FAULT NAME	DURATION
PS	Power supply	3 [h] 47 [min] 34[s]
PS	RF	3 [h] 13 [min] 33[s]

4	23:42	SUP	Global Post Mortem Event Event Timestamp: 10/06/12 23:42:39.163 Fill Number: 2718 Accelerator / beam mode: PROTON PHYSICS / STABLE BEAMS Energy: 4000080 [MeV] Intensity B1/B2: 15509 / 14217 [e^10 charges] Event Category / Classification: PROGRAMMED_DUMP / MULTIPLE_SYSTEM_DUMP First BIC input Triggered: First USR_PERMIT change: Ch 1-Programable Dump b1: A T -> F on CIB.CCR.LHC.B1
5	23:42	SUP	Global Post Mortem Event Confirmation Dump Classification: Programmed Dump Operator / Comment: papotti / End of physics fill, clean dump.
6	23:42	SUP	**BEAM MODE > BEAM DUMP** LHC RUN CTRL: BEAM MODE changed to BEAM DUMP
7	23:42	SUP	**BEAM MODE > BEAM DUMP** LHC RUN CTRL: BEAM MODE changed to BEAM DUMP
8	23:42	SUP	ELOGBOOK: STARTING B1 MKISS
9	23:43	SUP	ELOGBOOK: STARTING B2 MKISS
10	23:44	SUP	LHC SEQ: beam dump handshake closed; LHC=STANBY, EXP=VETO
11	23:44	SUP	LHC SEQ: MCS checks finished
12	23:45	SUP	LHC SEQ: SMP pre-operational checks finished
13	23:45	SUP	LHC SEQ: BIS pre-operational checks finished
14	23:48	SUP	**BEAM MODE > RAMP DOWN** LHC RUN CTRL: BEAM MODE changed to RAMP DOWN
15	23:48	SUP	LHC SEQ: BPMLHC calibration finished. Overall result: SUCCESS Chosen bunch spacing: (B1 & B2) BUNCH_50NSEC (manually chosen) (For more details see BI-LHC ELogBook)

Endeca Use Case

- Electronic Logbook
  - Endeca PoC

DEMO