Oracle Big Data Discovery for CERN's Control Data

Antonio Romero Marin

What's CERN

What's CERN

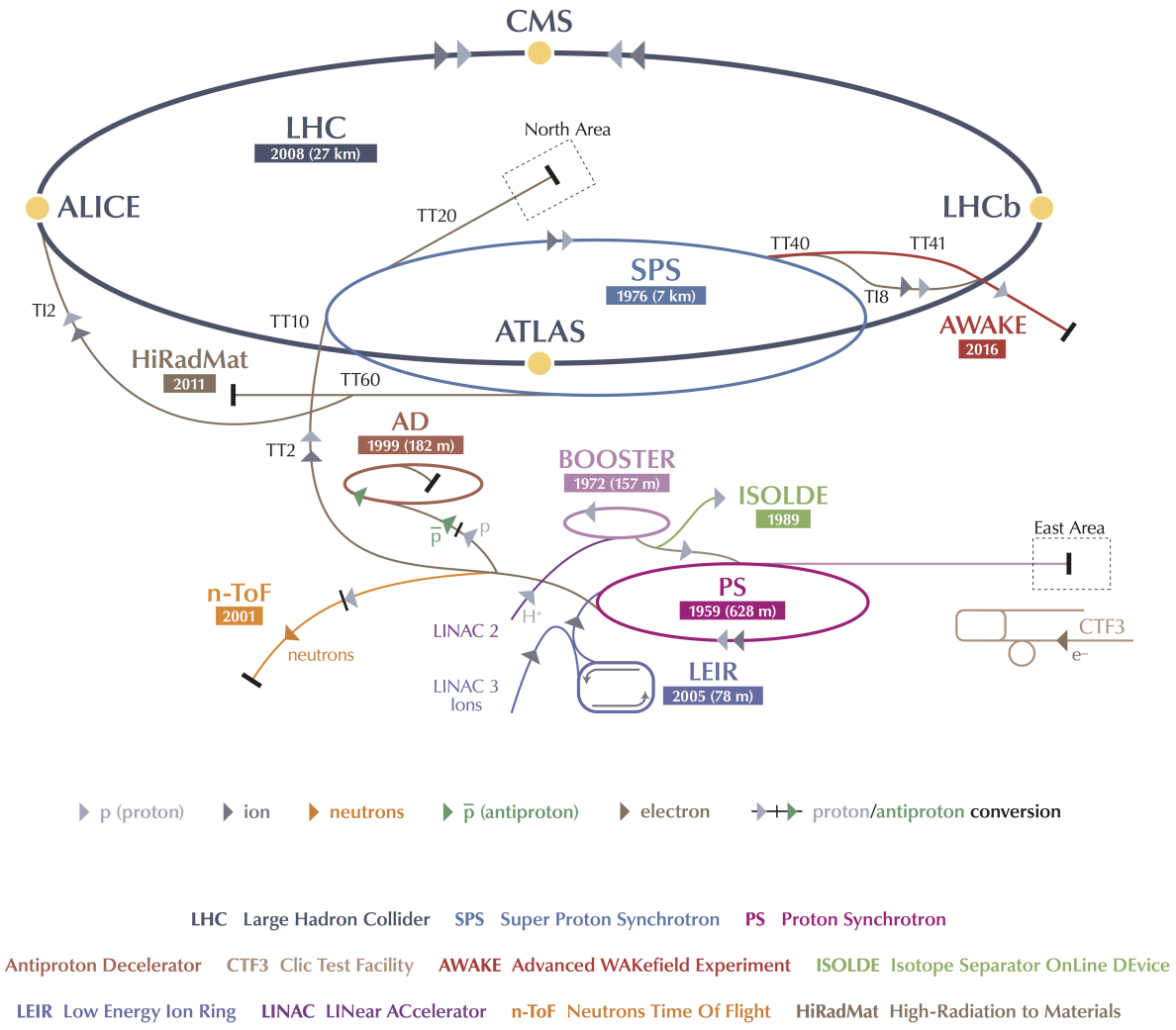
- European Laboratory for Particle Physics
  - Fundamental Research
- Worldwide International Collaboration
- Education & Training
- Push Frontiers of Technology

CERN openlab

- Public-private partnership between CERN and leading ICT companies and research institutes
- Accelerate cutting-edge solutions for the worldwide LHC community and wider scientific research.
- Designed to create and disseminate knowledge
  - Publication of reports and articles
  - Workshops or seminars
  - CERN openlab Student Programme

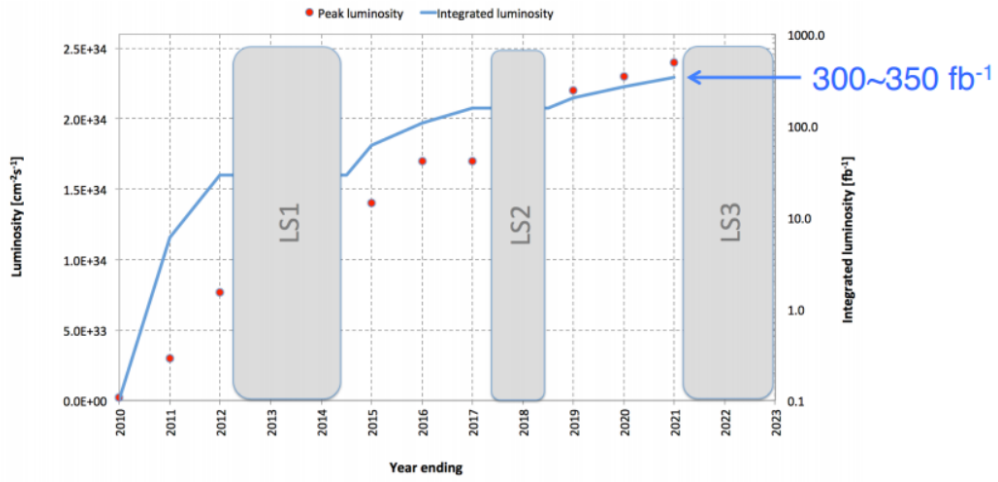
CERN Accelerator Complex

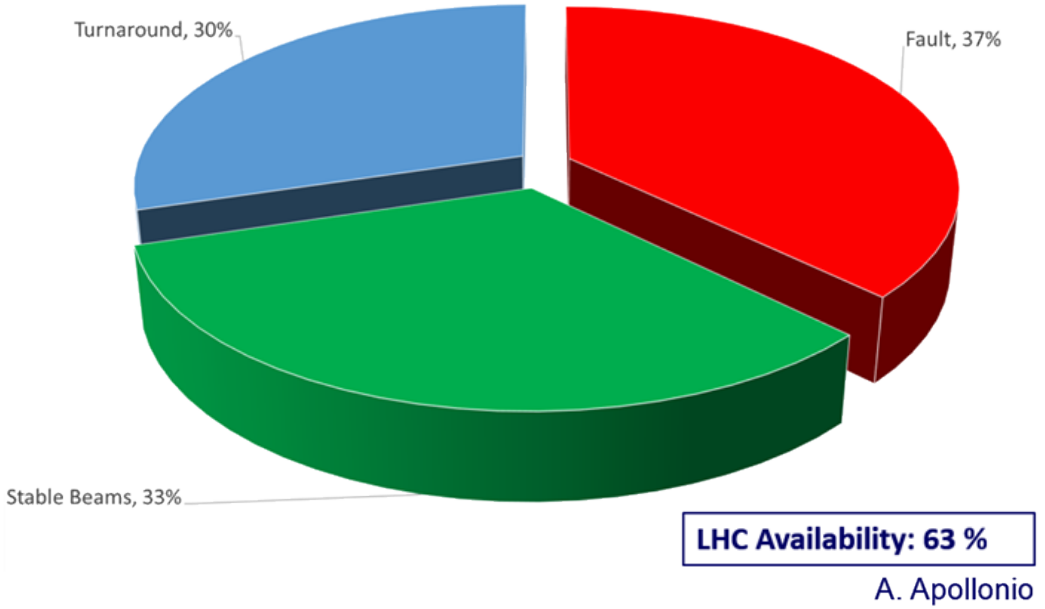
- Control and operations
  - Millions of sensors, signals
  - Large number of control devices
  - Equipment
- Monitoring and logging
- Supporting IT infrastructure
  - Databases
  - Network
  - Services

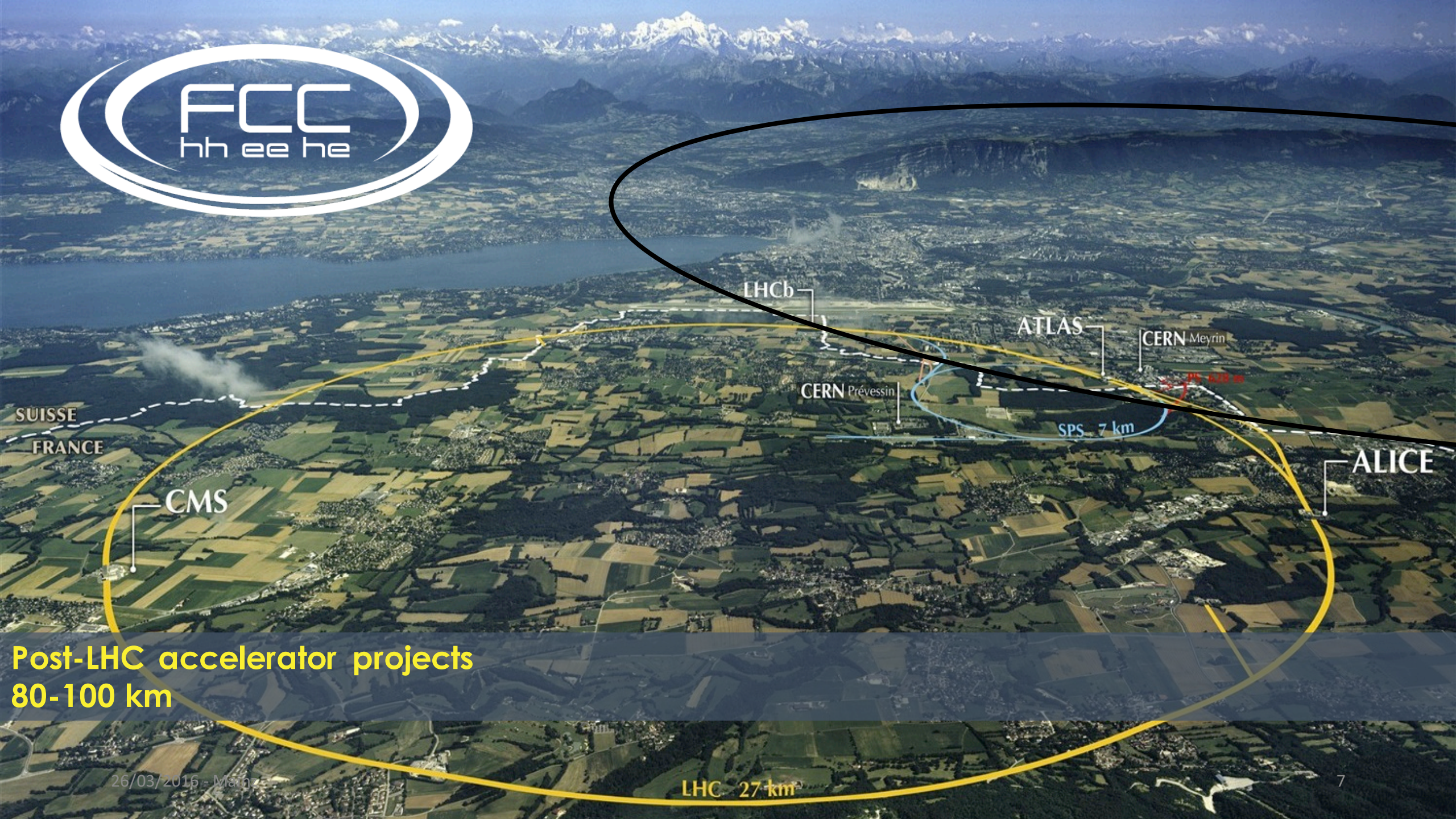


Data Analytics Challenges

- Some faults cannot be avoided
  - Decrease the availability for running physics
- Preventive maintenance is not enough
  - Does not take into account the condition of the equipment
- LHC upgrades will further increase luminosity
  - Computing resources needs will be higher
  - Data generated will increase drastically

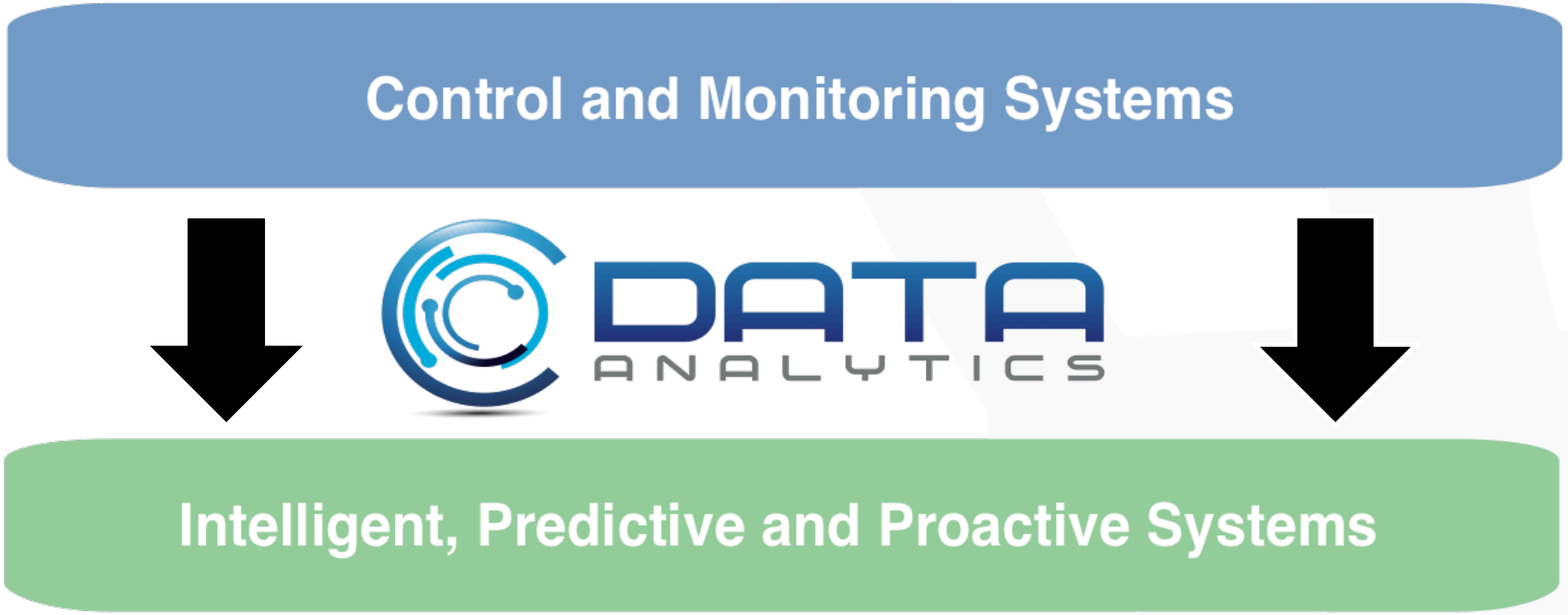






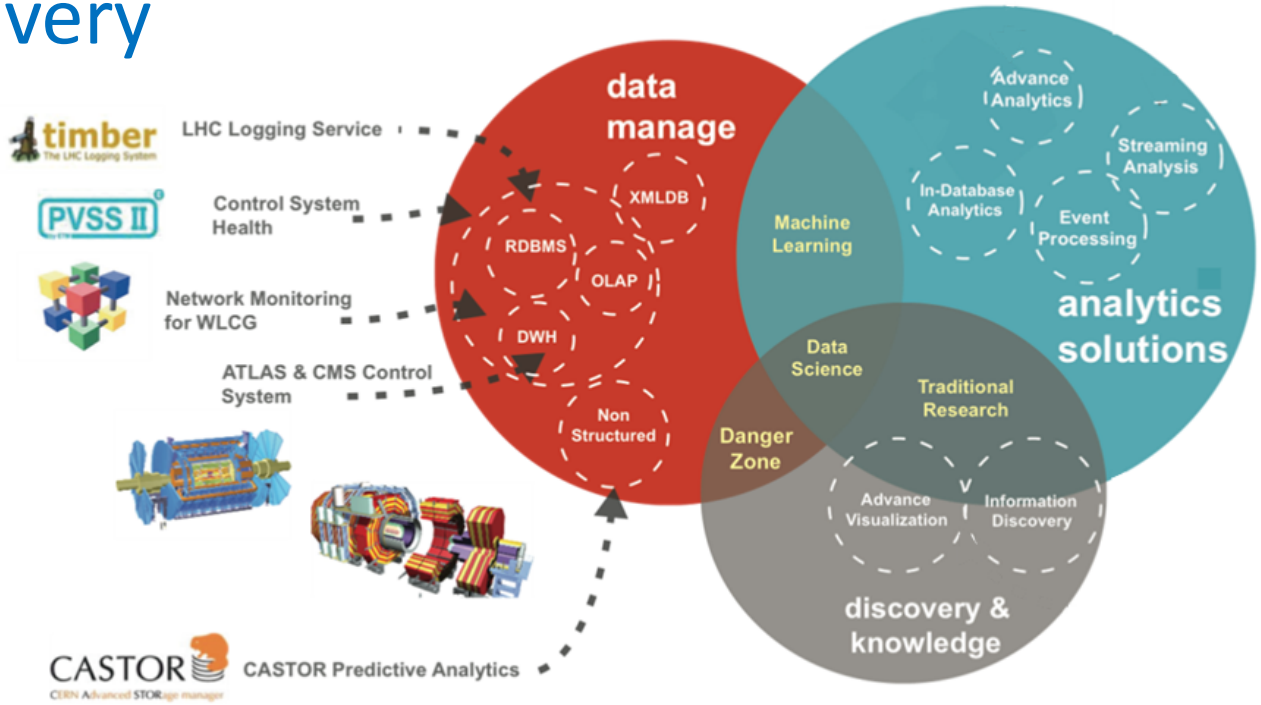
Post-LHC accelerator projects
80-100 km

Data Analytics Objective



Areas of investigation

- Predictive maintenance and system optimization
- Data extraction, transformation and loading (ETL)
- Data Visualization and Discovery



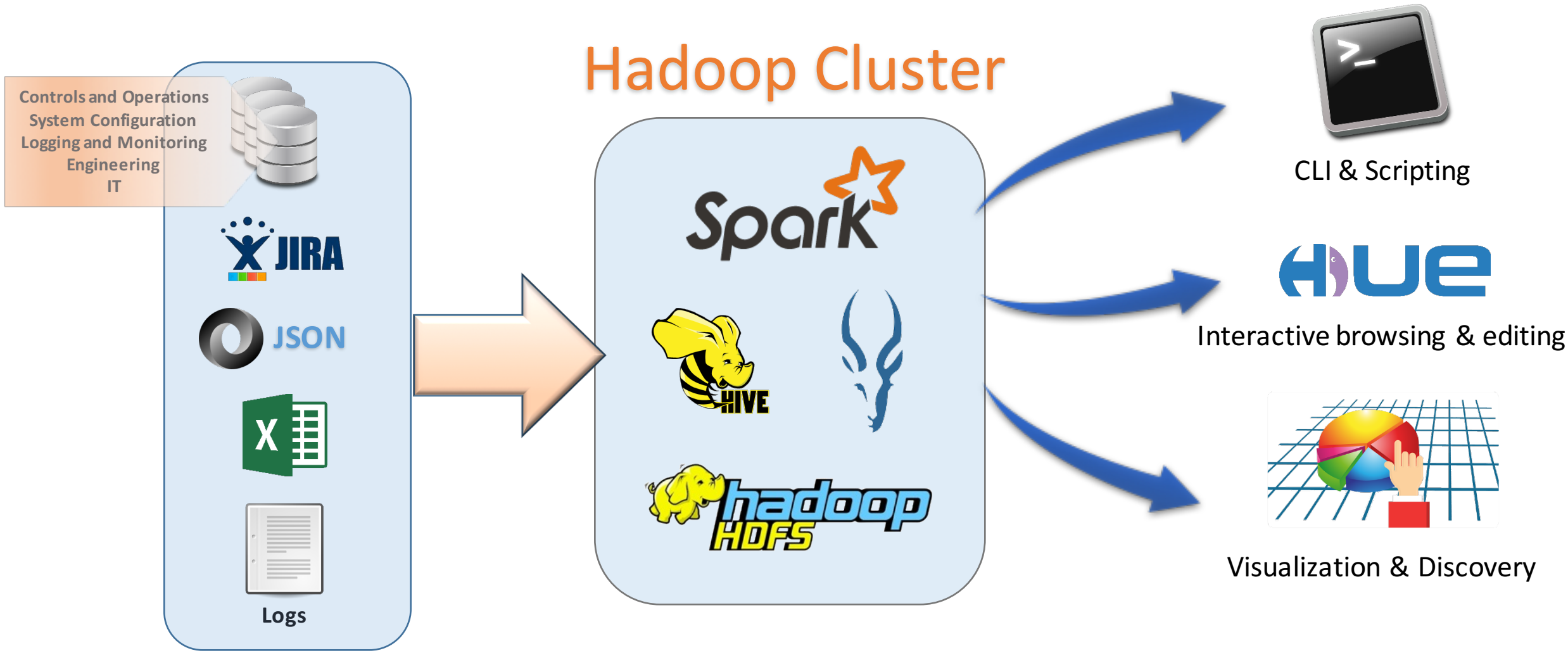
Use Case - FCC RAMS studies

- Reliability, Availability, Maintainability and Safety (RAMS) studies for the Future Circular Collider (FCC)
- Study and increase the reliability and availability of the LHC
- Use RAMS findings to assess the feasibility of the needs of FCC
- Data distributed across multiple sources
  - Operations e-logbook
  - Accelerator Fault Tracking project
  - Accelerator logging service
  - Accelerator schedules
  - Cryogenics
  - Add more…
    - Vacuum, Power Converters, etc.

Requirements

- Flexible
  - Data => structured, semi-structured and non-structured, data editing
  - Use => Interactive, CLI & Scripting
- ETL functionalities
- Scalable
  - Data is foreseen to increase significantly (+datasets)
  - Processing
- Powerful
  - Browse data
  - Correlate Information
  - Visualization
  - Analytics

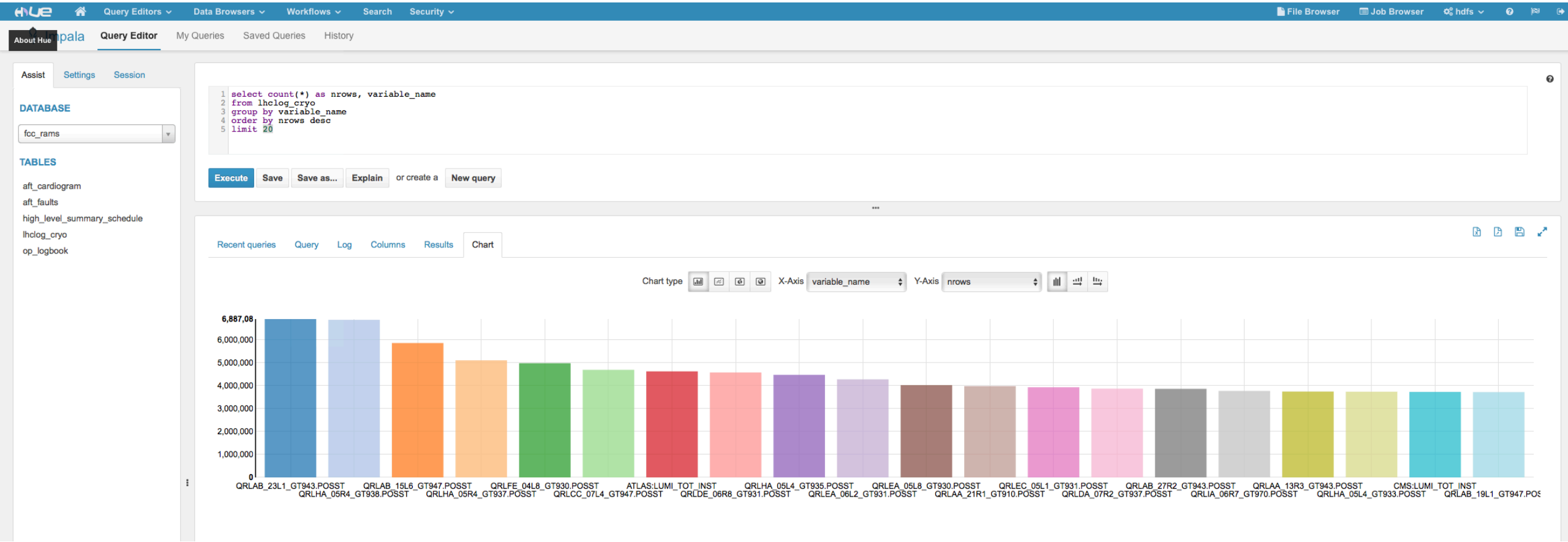
Hadoop based solution



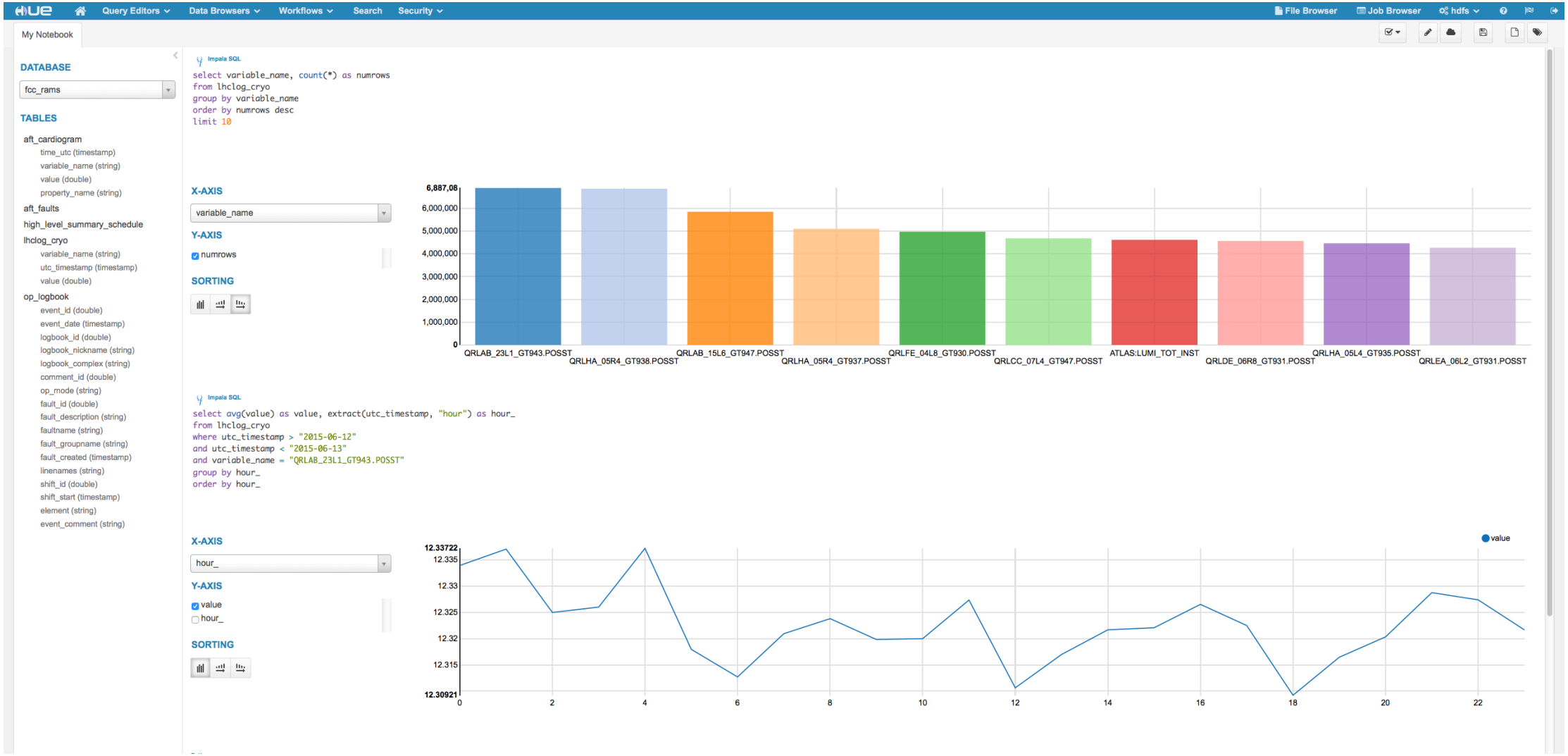
HUE – Hadoop User Experience

- Hue is an open source suite of web-based applications for analyzing data with any Apache Hadoop
- It features:
  - SQL Editors for Hive, Impala, MySql, PostGres, Sqlite and Oracle
  - Dynamic search dashboards for Solr
  - Spark Notebooks
  - Browsers for YARN, HDFS, Hive table Metastore, HBase, ZooKeeper
  - Pig Editor, Sqoop2, Oozie workflows Editors and Dashboards
  - Wizards to import data into Hadoop

HUE – Hadoop User Experience



HUE – Hadoop User Experience



HUE – Hadoop User Experience

- One tool to use multiple Hadoop components
- Easy to use
- Compatible with multiple versions, open source
- Extensible
- But
  - Requires language knowledge to explore and transform the data
  - Limited for Data Discovery

Data Discovery

- Interactive and visual analytics
  - Find hidden patterns
  - Get **new insights**
- 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
- Integrate multiple data sources
  - Analyze information of any type and any source

Oracle Big Data Discovery Overview

- Data Exploration & Discovery
  - Interactive catalog of all data
  - Assess attribute statistics, data quality and outliers
  - Quick data exploration or create dashboards and applications
- Data Transformation with Spark in Hadoop
  - Apply built-in transformations or write your own scripts
  - Data Enrichment
    - Text: Entity extraction, relevant terms, sentiment, language detection
    - Geographical information: address, IP, reverse
  - Preview results, undo, commit and replay transforms
- Collaborative environment
  - Share and bookmarks
  - Create and share transformed datasets

Components

- BDD Data Processing (Spark on YARN)
  - Hive Table Detector
  - Profiling and sampling
  - Transformations and enrichments
  - Refresh/incremental update datasets automatically or manually
- Dgraph (In-Memory Discovery Indexes)
  - In-memory, columnar, multi-core architecture
- Web Studio
  - Catalog, explore, transform and discover UI's

Architecture overview

cloudera

CDH 5.5.1
16 nodes, 24 GB ram
Intel Xeon L5520 @ 2.27GHz
165 TB HDFS

Coordination

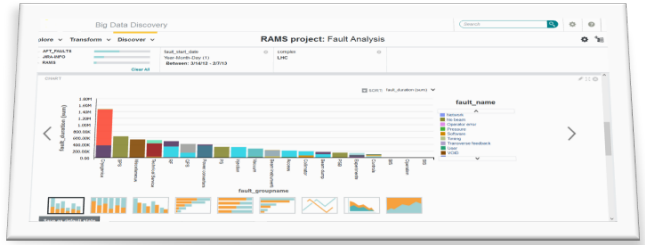
Oracle Big Data Discovery
Libraries + Hive table detector

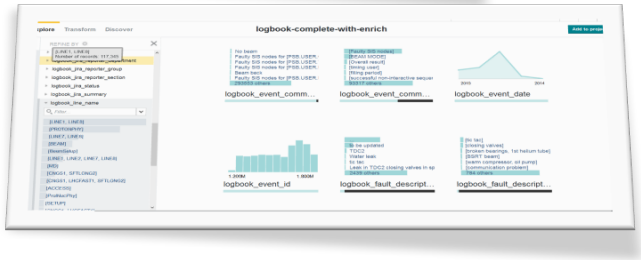
HIVE | Spark

Resource Management (YARN)

Data Storage — hadoop HDFS

Data Integration — Oozie | FLUME



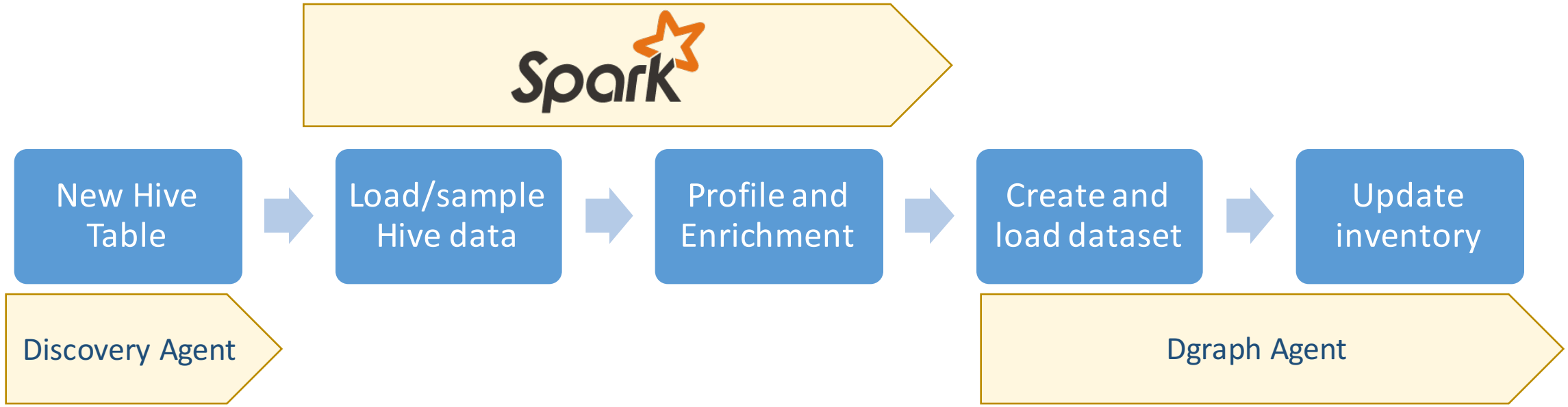


Dgraph
In-Memory Columnar
Database

ORACLE
EXALYTICS

4x Xeon E7-8895 v2 (15 cores each)
2 TB RAM
4.8 TB Flash + 6 x 1.2 TB 10K HDD

Data Processing Workflow

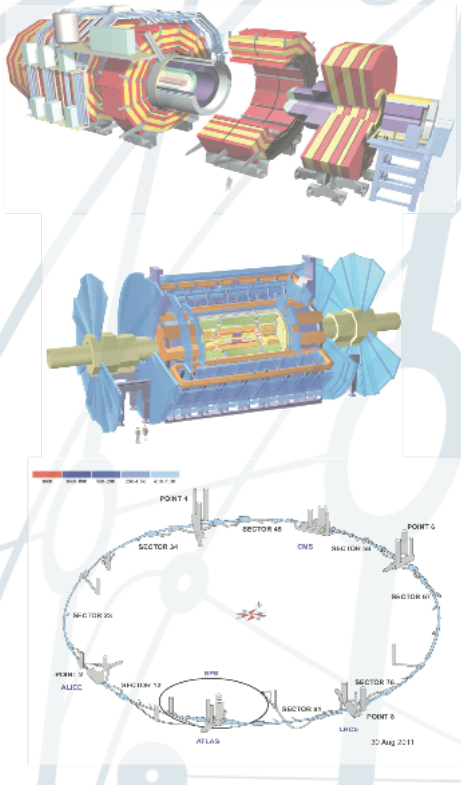


Important Technical Features

- Supports HDP and CDH Hadoop distributions
- Kerberos support
- Spark on YARN
- Data set and project level security in Studio
- Incremental updates and refresh datasets
- Applications and dashboards
- Ability to publish and share transformation scripts
- Custom visualization support (Javascript, D3, EQL)

Use Case: FCC RAMS

Scenario



LogBook

MEDB

CALS

AFT

AME

- Power Converters
- Cryogenics
- Machine Protection
- Accelerator Major Events
- Accelerator Fault Tracking
- Accelerator Logging
- Operations logbook

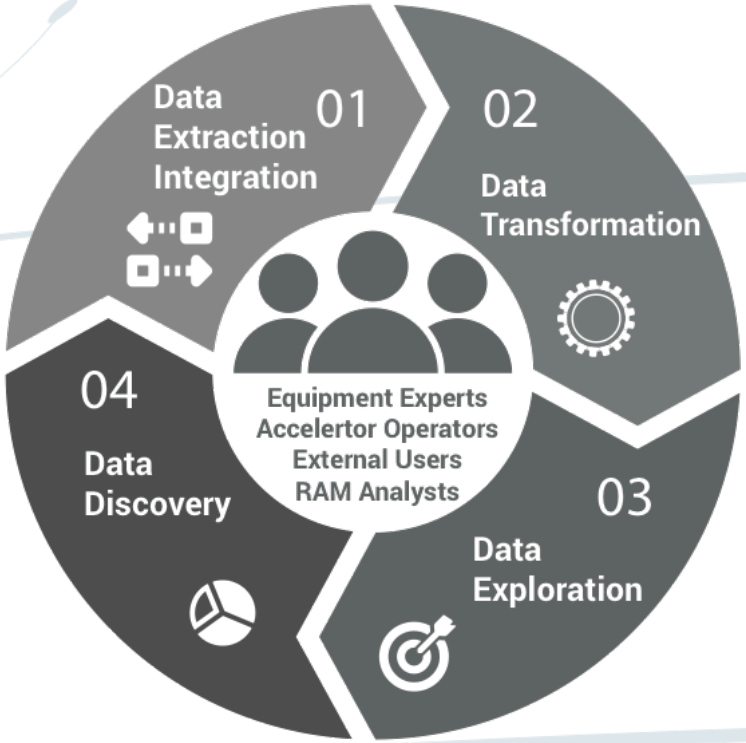
NoSQL

XML

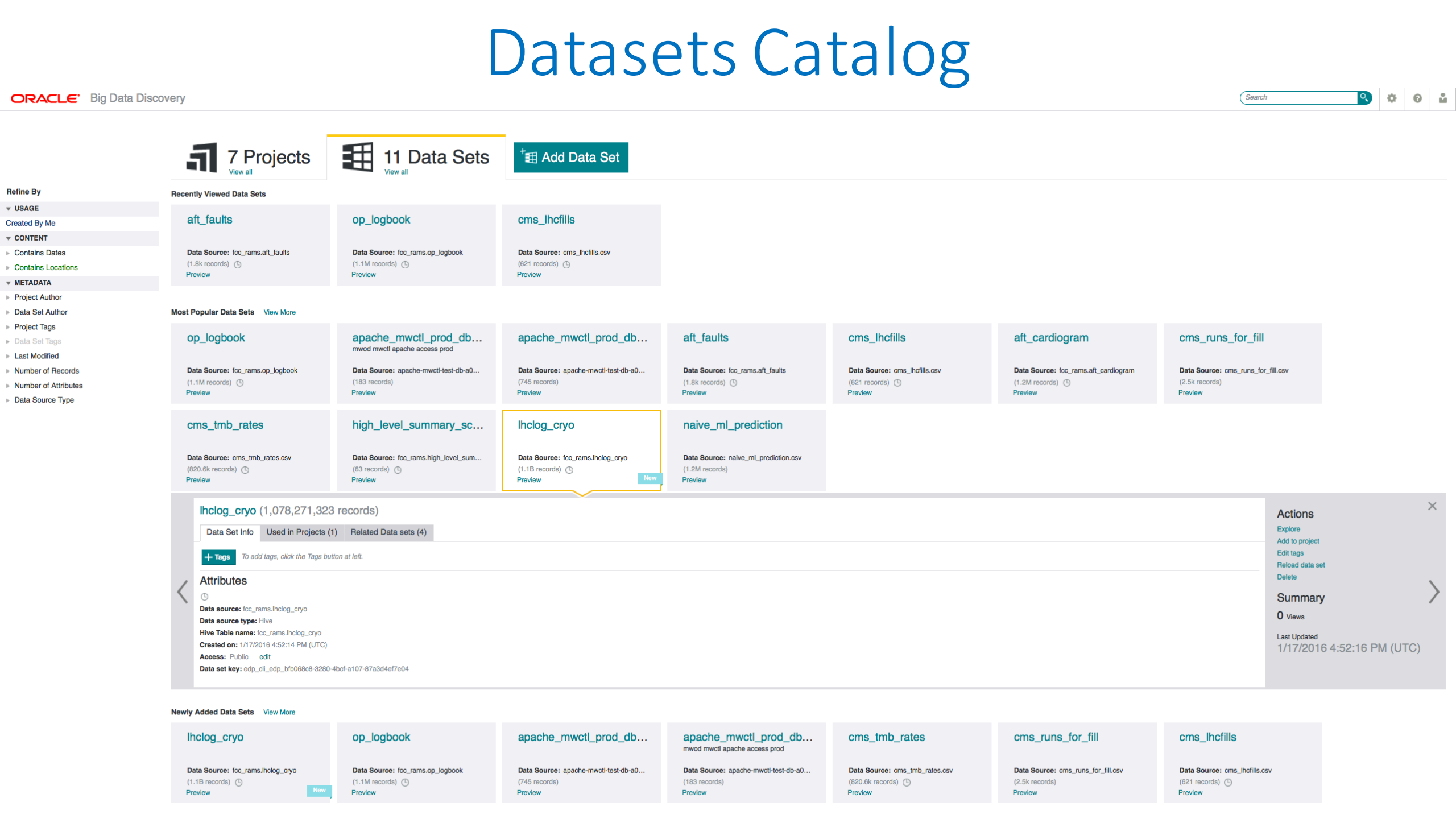
JSON

Text

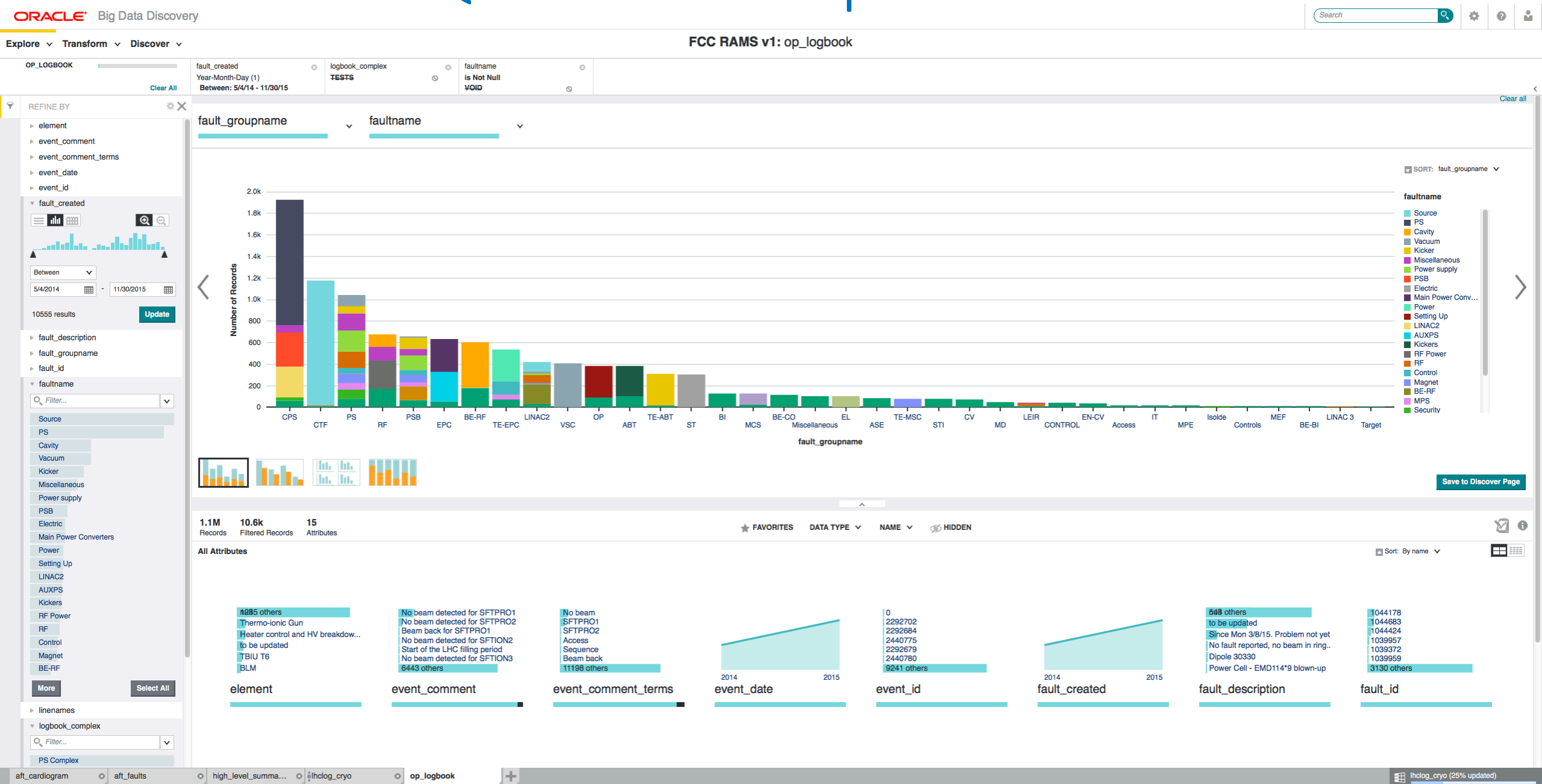
RDBMS



Datasets Catalog



Quick Data Exploration



Data Transformation UI - ETL

ORACLE Big Data Discovery

Explore Transform Discover

FCC RAMS v1: aft_faults

Transformation Editor

Use refinement state as a conditional statement

Enable automatic typeahead + Functions + Attribute

```
1  diffDates(op_end_time, creation_time_utc, SECONDS)
```

Configure Output Settings

Apply to "op_duration"

Create New Attribute

New Attribute Name: op_duration_seconds

Data Type: Double

Single Assign

Cancel Preview Add to Script

TRANSFORM SCRIPT

elogbook_fault_id - Transform

fault_id - Create

fault_classification_id - Transform

Commit to Project

1.8k Records **1.8k** Filtered Records **21** Attributes

FAVORITES DATA TYPE NAME HIDDEN

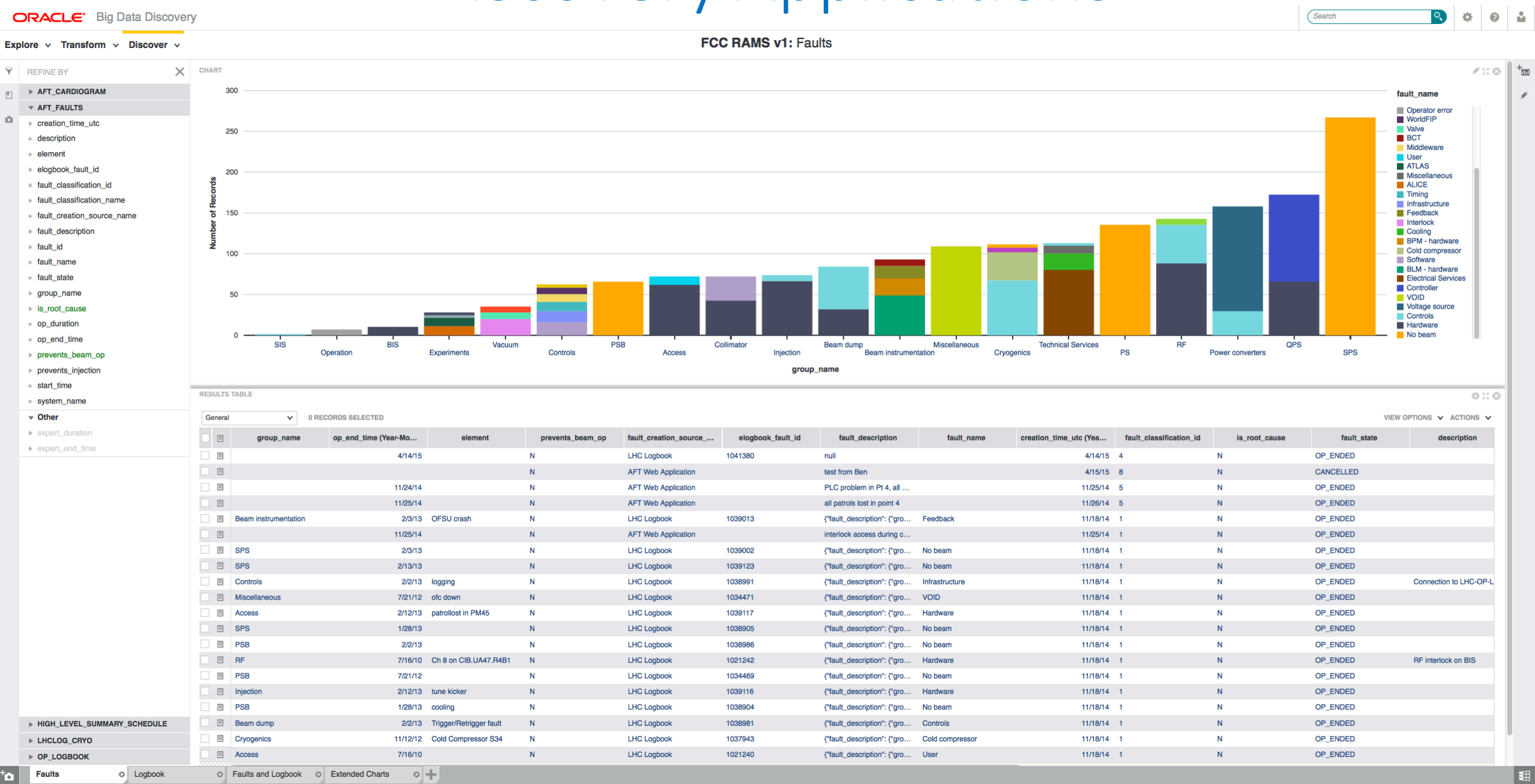
All Attributes

Sort: By preview order

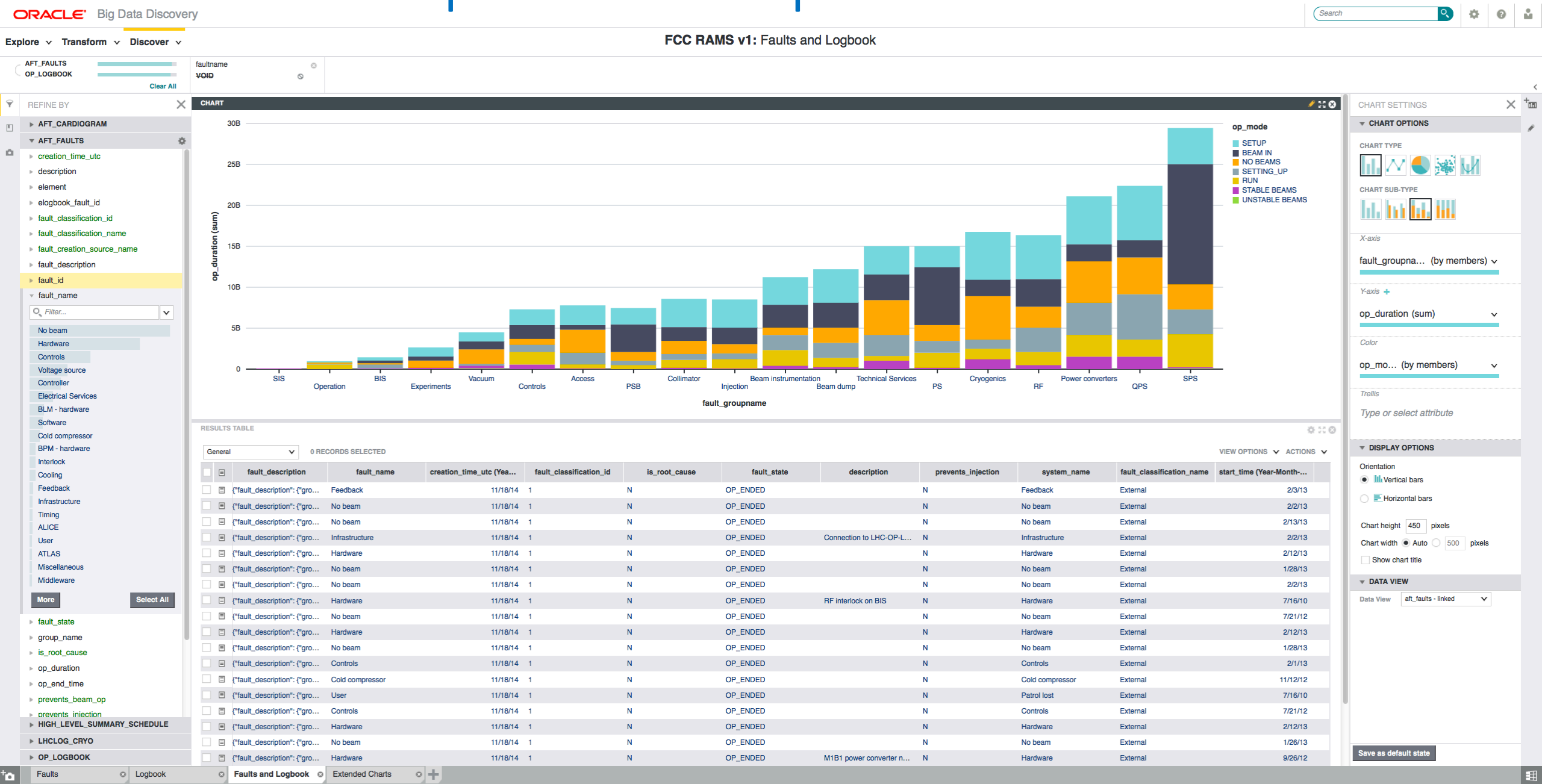
	fault_id	fault_name	fault_state	group_name	is_root_cause	op_duration	op_duration_seconds	op_end_time	prevents_beam_op	prevents_injection	start_time
	15160		OP_ENDED		N	4,266	10,932	2015-04-14 08:28:38 UTC	N	Y	2015-04-14 07:17:32 UTC
	15223		CANCELLED		N				N	N	2015-04-15 06:30:00 UTC
ll secto…	3761		OP_ENDED		N	21,588	47,292	2014-11-24 17:30:38 UTC	N	N	2014-11-24 11:30:50 UTC
	3765		OP_ENDED		N	99,735	66,405	2014-11-25 15:17:19 UTC	N	N	2014-11-24 11:35:04 UTC
roup_…	1909	Feedback	OP_ENDED	Beam instrumentation	N	66	56,455,621	2013-02-03 04:20:39 UTC	N	N	2013-02-03 04:19:33 UTC
condit…	3781		OP_ENDED		N	78,786	-9,846	2014-11-25 09:23:53 UTC	N	N	2014-11-24 11:30:47 UTC
roup_…	1908	No beam	OP_ENDED	SPS	N	20,870	56,457,295	2013-02-03 03:52:45 UTC	N	N	2013-02-02 22:04:55 UTC
roup_…	1924	No beam	OP_ENDED	SPS	N	147	55,550,285	2013-02-13 15:49:35 UTC	N	N	2013-02-13 15:47:08 UTC
roup_…	1907	Infrastructure	OP_ENDED	Controls	N	6,760	56,538,706	2013-02-02 05:15:54 UTC	N	N	2013-02-02 03:23:14 UTC
roup_…	1471	VOID	OP_ENDED	Miscellaneous	N	2,614	73,416,728	2012-07-21 20:55:32 UTC	N	N	2012-07-21 20:11:58 UTC
roup_…	1923	Hardware	OP_ENDED	Access	N	3,985	55,624,990	2013-02-12 19:04:30 UTC	N	N	2013-02-12 17:58:05 UTC
roup_…	1889	No beam	OP_ENDED	SPS	N	4,292	56,956,446	2013-01-28 09:13:34 UTC	N	N	2013-01-28 08:02:02 UTC
roup_…	1906	No beam	OP_ENDED	PSB	N	8,765	56,548,197	2013-02-02 02:37:43 UTC	N	N	2013-02-02 00:11:38 UTC
roup_…	117	Hardware	OP_ENDED	RF	N	3,750	137,030,838	2010-07-16 14:20:22 UTC	N	N	2010-07-16 13:17:52 UTC
roup_…	1470	No beam	OP_ENDED	PSB	N	1,322	73,419,672	2012-07-21 20:06:28 UTC	N	N	2012-07-21 19:44:26 UTC
roup_…	1922	Hardware	OP_ENDED	Injection	N	14,065	55,629,441	2013-02-12 17:50:19 UTC	N	N	2013-02-12 13:55:54 UTC
roup_…	1888	No beam	OP_ENDED	PSB	N	7,090	56,970,250	2013-01-28 05:23:30 UTC	N	N	2013-01-28 03:25:20 UTC
roup_…	1905	Controls	OP_ENDED	Beam dump	N	2,773	56,556,800	2013-02-02 00:14:20 UTC	N	N	2013-02-01 23:28:07 UTC
roup_…	1774	Cold compressor	OP_ENDED	Cryogenics	N	3,307	63,595,623	2012-11-12 13:00:37 UTC	N	N	2012-11-12 12:05:30 UTC
roup_…	116	User	OP_ENDED	Access	N	5,859	137,044,782	2010-07-16 10:27:58 UTC	N	N	2010-07-16 08:50:19 UTC
roup_…	1469	Controls	OP_ENDED	Beam dump	N	4,625	73,421,636	2012-07-21 19:33:44 UTC	N	N	2012-07-21 18:16:39 UTC
roup_…	1921	Hardware	OP_ENDED	RF	N	20,968	55,624,990	2013-02-12 19:04:30 UTC	N	N	2013-02-12 13:15:02 UTC
roup_…	1887	No beam	OP_ENDED	SPS	N	2,107	57,082,801	2013-01-26 22:07:39 UTC	N	N	2013-01-26 21:32:32 UTC
roup_…	1633	Hardware	OP_ENDED	RF	N	6,148	67,696,057	2012-09-26 02:00:03 UTC	N	N	2012-09-26 00:17:35 UTC
roup_…	1904	No beam	OP_ENDED	SPS	N	318	56,594,555	2013-02-01 13:45:05 UTC	N	N	2013-02-01 13:39:47 UTC
roup_…	1773	Hardware	OP_ENDED	RF	N	10,413	63,601,399	2012-11-12 11:24:21 UTC	N	N	2012-11-12 08:30:48 UTC
roup_…	115	Controls	OP_ENDED	Power converters	N	3,815	137,052,743	2010-07-16 08:15:17 UTC	N	N	2010-07-16 07:11:42 UTC
roup_…	1468	No beam	OP_ENDED	SPS	N	2,733	73,441,164	2012-07-21 14:08:16 UTC	N	N	2012-07-21 13:22:43 UTC

aft_cardiogram | aft_faults | high_level_summa… | lhclog_cryo | op_logbook

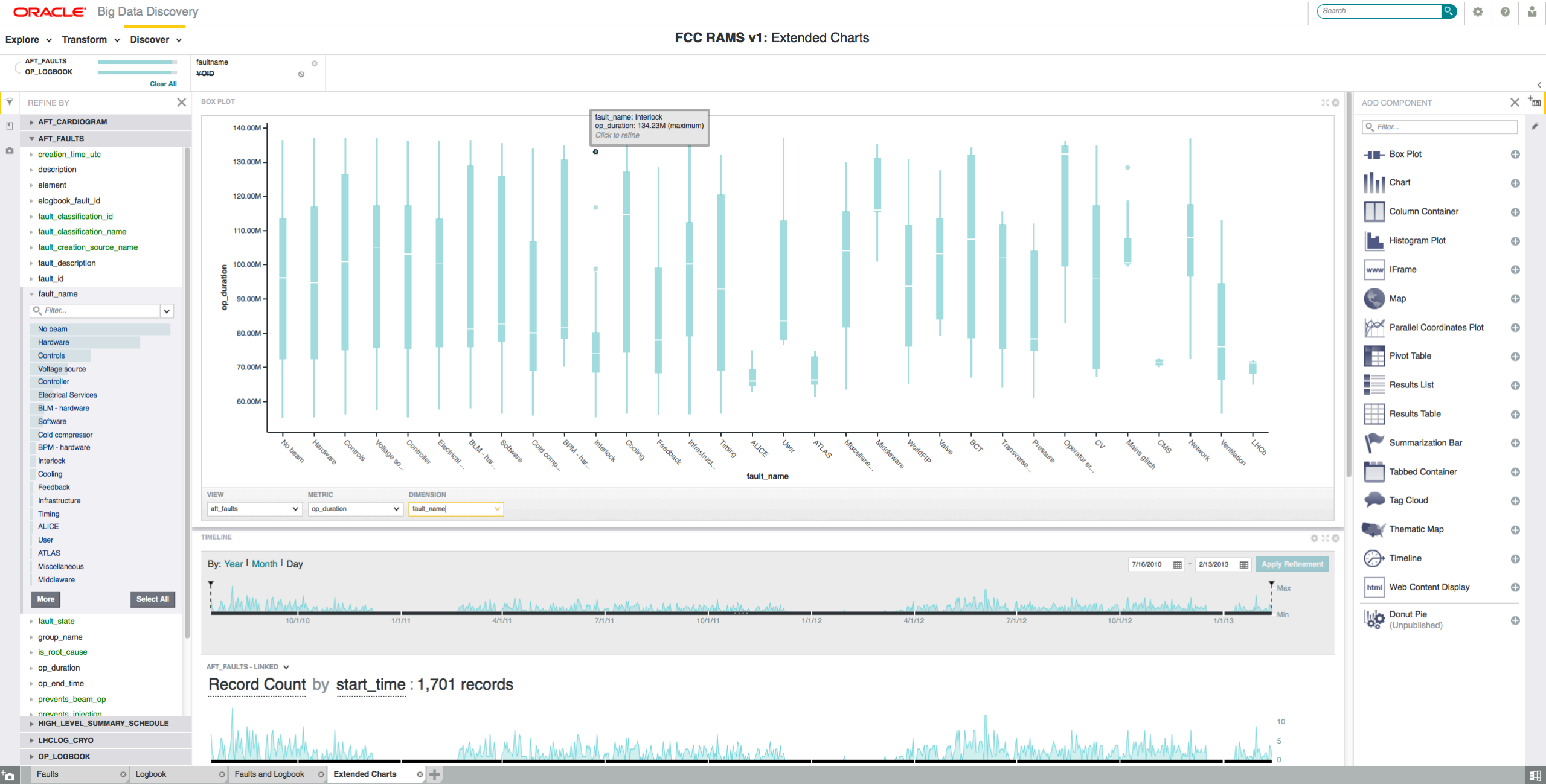
Discovery Applications



Explore Multiple Datasets



Data Visualization



Data Visualization

ORACLE Big Data Discovery

Search

Explore Transform Discover

CMS Fills: tmb_rates

CMS_RUNS_FOR_...
CMS_TMB_RATES
Clear All

runnumber
199608

REFINE BY

- CMS_LHCFILLS
- CMS_RUNS_FOR_FILL
- CMS_TMB_RATES
  - 1/1
  - 1/2
  - 1/3
  - 2/1
  - 2/2
  - 3/1
  - 3/2
  - 4/1
  - 4/2
  - fct_time
  - instlumi
  - instlumierror
  - me
  - pileup
  - pileuprms
  - runnumber

Filter...

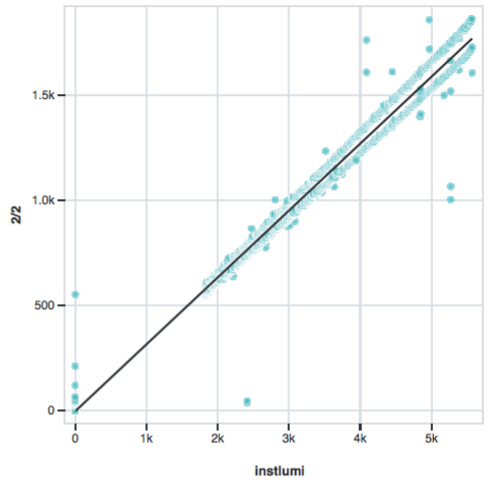
252504
255038
173692
177730
258177
207454
200091
206745
166512
194050
195552
166380
195398
172822
258158
165993
193621
210658
202504
176309

More Select All

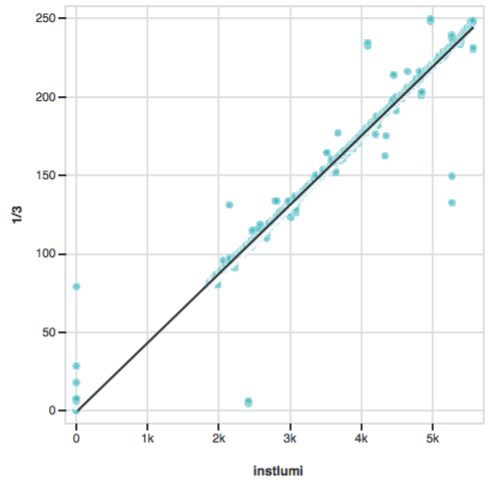
- starttime
- stoptime

COLUMN CONTAINER

CHART



CHART



RESULTS TABLE

0 RECORDS SELECTED

VIEW OPTIONS ACTIONS

runnumber	1/1	1/2	1/3	2/1	2/2	3/1	3/2	4/1	4/2	fct_time (Year-Month-D...	instlumi	instlumierror
199608	67,350.85	2,642.26	206.71	31,179.86	1,442.91	15,986.39	1,996.61	14,363.54	27,744.87	7/24/12	4,712.53	0.
199608	67,365.72	2,642.37	207.29	31,161.58	1,443.19	15,975.94	1,994.15	14,358.52	27,796.3	7/24/12	4,715.04	0.
199608	67,454.42	2,643.77	207.7	31,204.16	1,442.88	16,002.51	1,997.64	14,372.98	27,779.07	7/24/12	4,721.7	0.
199608	67,509.79	2,646.2	207.54	31,248.23	1,443.6	16,016.13	1,999.95	14,384.57	27,841.57	7/24/12	4,722.9	0.
199608	68,388.87	2,682.23	210.2	31,660.51	1,463.29	16,227.77	2,025.69	14,576.27	28,205.57	7/24/12	4,785.57	0.
199608	67,534.9	2,650.59	208.6	31,258.85	1,447.08	16,030.38	2,001.05	14,396.56	27,861.61	7/24/12	4,726.93	0.
199608	68,442.37	2,685.5	210.66	31,672.86	1,466.04	16,229.6	2,028.39	14,572.22	28,179.21	7/24/12	4,787.21	0.
199608	57,447.28	2,242.93	175.22	26,464.05	1,223.52	13,601.92	1,693.48	12,218.61	23,482.16	7/24/12	4,000.56	0.
199608	67,631.33	2,649.78	208.45	31,308.91	1,446.96	16,056.75	2,002.4	14,426.47	27,878.9	7/24/12	4,729.61	0.
199608	68,650.39	2,690.54	210.96	31,721.55	1,467.61	16,260.08	2,031.54	14,605.17	28,235.37	7/24/12	4,790.64	0.
199608	41,589.1	1,623.45	126.95	19,276.53	881.4	9,906.8	1,229.46	8,899.11	16,878.53	7/25/12	2,912.43	0.
199608	57,490.17	2,240.56	175.86	26,474.51	1,225.51	13,605.17	1,693.86	12,224.77	23,468.41	7/24/12	4,004.74	0.
199608	67,674.77	2,652.31	208.49	31,322.94	1,449.79	16,055.61	2,006.47	14,424.57	27,900.23	7/24/12	4,734.54	0.
199608	70,365.46	2,743.45	215.43	32,362.59	1,497.95	16,585.46	2,070.38	14,905.43	28,829.61	7/24/12	4,894.94	0.
199608	69,031.09	2,686.93	211.71	31,739.6	1,466.16	16,273.74	2,029.41	14,625.27	28,267.1	7/24/12	4,798.35	0.
199608	150.47	36.83	3.62	314.42	29.67	289.97	42.38	261.89	58.85	7/24/12	0.32	
199608	28,283	1,166.89	91.61	14,532.89	687.88	7,316.02	885.32	6,180.48		7/25/12	2,088.17	0.

Fills | Runs For Fill | tmb_rates

Next steps

- Check new upcoming features presented during OOW 15
  - Scalability improvements
  - Streaming integration (Kafka)
  - Add more advanced charts and visualizations
  - Machine Learning integration
- Evaluate Big Data Discovery cloud
- Extend to more CERN use cases
  - Controls and Operations
    - Accelerator Fault Tracking
    - Diagnostics and Monitoring
  - IT Infrastructure Monitoring
    - Server logs analysis
    - Database latency
  - Human Resources

Conclusions

- Data visualization and discovery is an important area in data analytics
  - Facilitates users to visualize and explore their data
  - Find correlations, extract insight and useful information
- Important points
  - Flexible and user-friendly platform
  - Advanced data visualization and exploration
  - Collaborative
- Application to different domains
  - Controls and Operations
  - IT Infrastructure Monitoring
  - Human Resources

www.cern.ch