
CERN

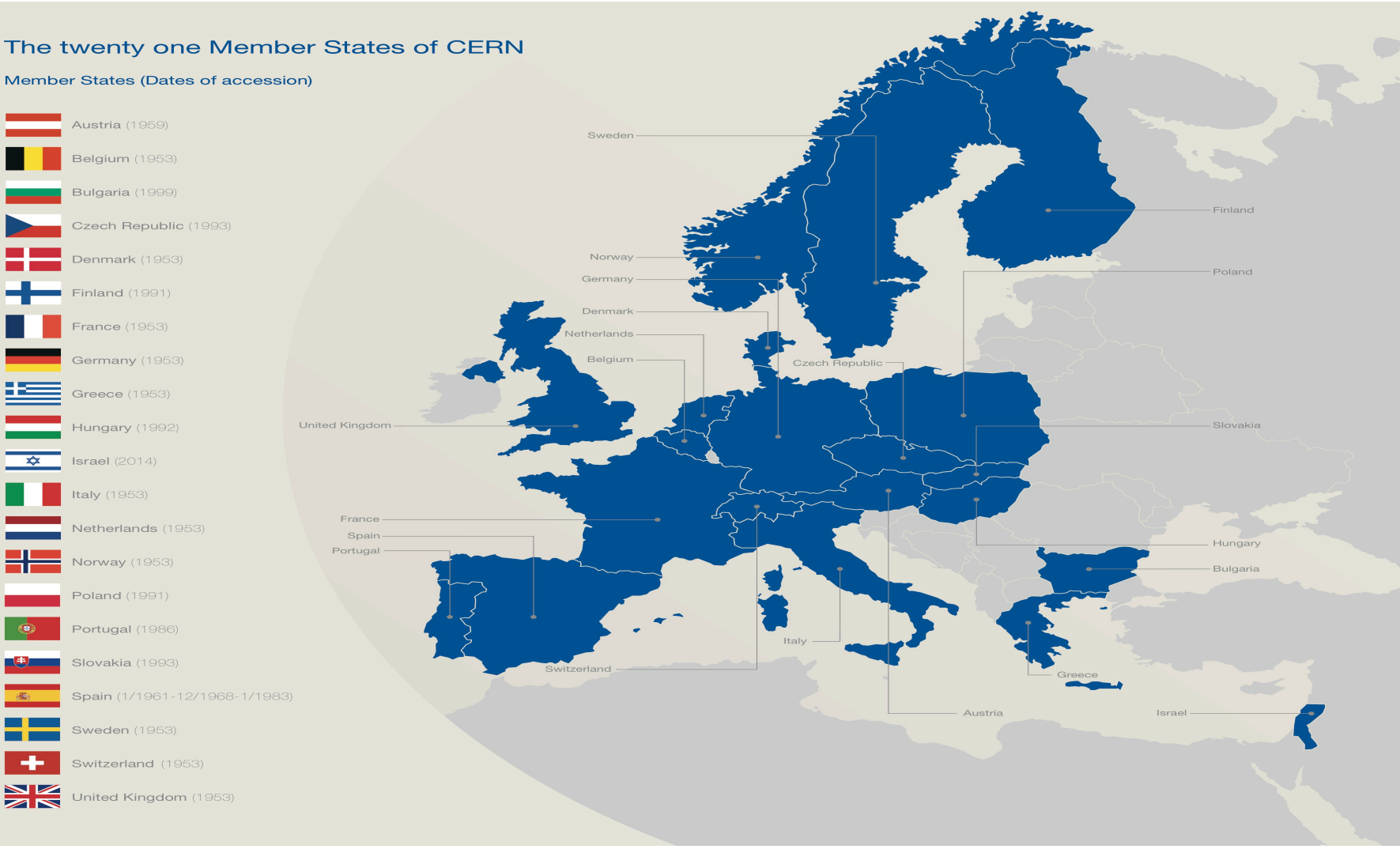
CERN's Journey into Big Data and Analytics

Manuel Martin Marquez
Senior Data Scientists and Data Streaming Project Leader
CERN – European Organization for Nuclear Research

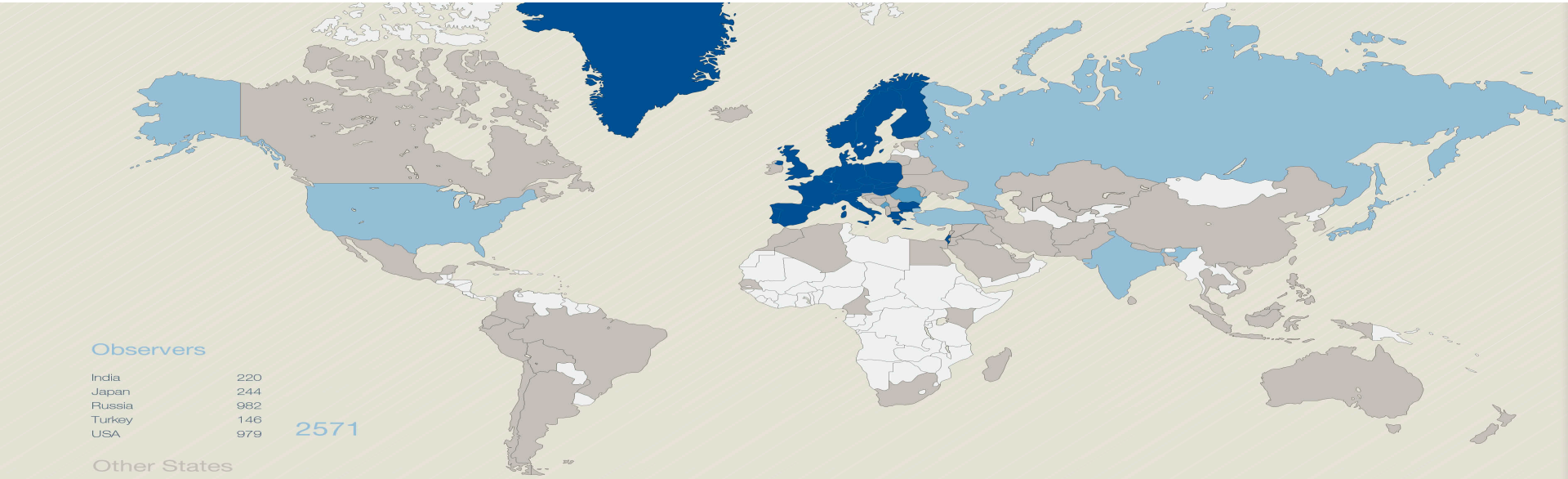
Agenda

1. CERN Introduction
2. Why CERN cares about Big Data?
3. What does Big Data looks like? What are the benefits and risks?
4. The data process: jumping into Big Data
5. Profiting from the data investments
6. Making the most of it: Advance and scalable analytics
7. Some conclusions from our experince.

01. CERN – European Organization for Nuclear Research



01. CERN – A Worldwide Collaboration



Observers

India	220
Japan	244
Russia	982
Turkey	146
USA	979

2571

Other States

Afghanistan	1	El Salvador	1	Pakistan	41
Albania	2	Estonia	16	Palestine (O.T.).	4
Algeria	8	Georgia	36	Peru	8
Argentina	11	Gibraltar	1	Philippines	1
Armenia	25	Hong Kong	1	Saudi Arabia	3
Australia	25	Iceland	4	Senegal	1
Azerbaijan	8	Indonesia	1	Singapore	2
Bangladesh	4	Iran	28	Sint Maarten	2
Belarus	47	Ireland	22	Slovenia	27
Bolivia	3	Jordan	2	South Africa	16
Bosnia & Herzegovina	1	Kenya	1	Sri Lanka	5
Brazil	108	Korea, D.P.R.	1	Syria	2
Cameroon	1	Korea Rep.	117	Thailand	12
Canada	134	Kuwait	1	T.F.Y.R.O.M.	1
Cape Verde	1	Lebanon	12	Tunisia	6
Chile	12	Lithuania	19	Ukraine	55
China	280	Luxembourg	4	Uzbekistan	4
China (Tapei)	45	Madagascar	4	Venezuela	9
Colombia	30	Malaysia	15	Viet Nam	9
Croatia	35	Mauritius	1	Zimbabwe	2
Cuba	7	Mexico	64		
Cyprus	16	Montenegro	3		
Ecuador	3	Morocco	12		
Egypt	19	Nepal	5		
		New Zealand	7		

1415

Member States

Austria	99	Greece	152	Slovakia	88
Belgium	106	Hungary	68	Spain	337
Bulgaria	75	Israel	51	Sweden	75
Czech Republic	202	Italy	1686	Switzerland	180
Denmark	53	Netherlands	153	United Kingdom	640
Finland	87	Norway	61		
France	751	Poland	229		
Germany	1150	Portugal	109		

6352

Candidate for Accession

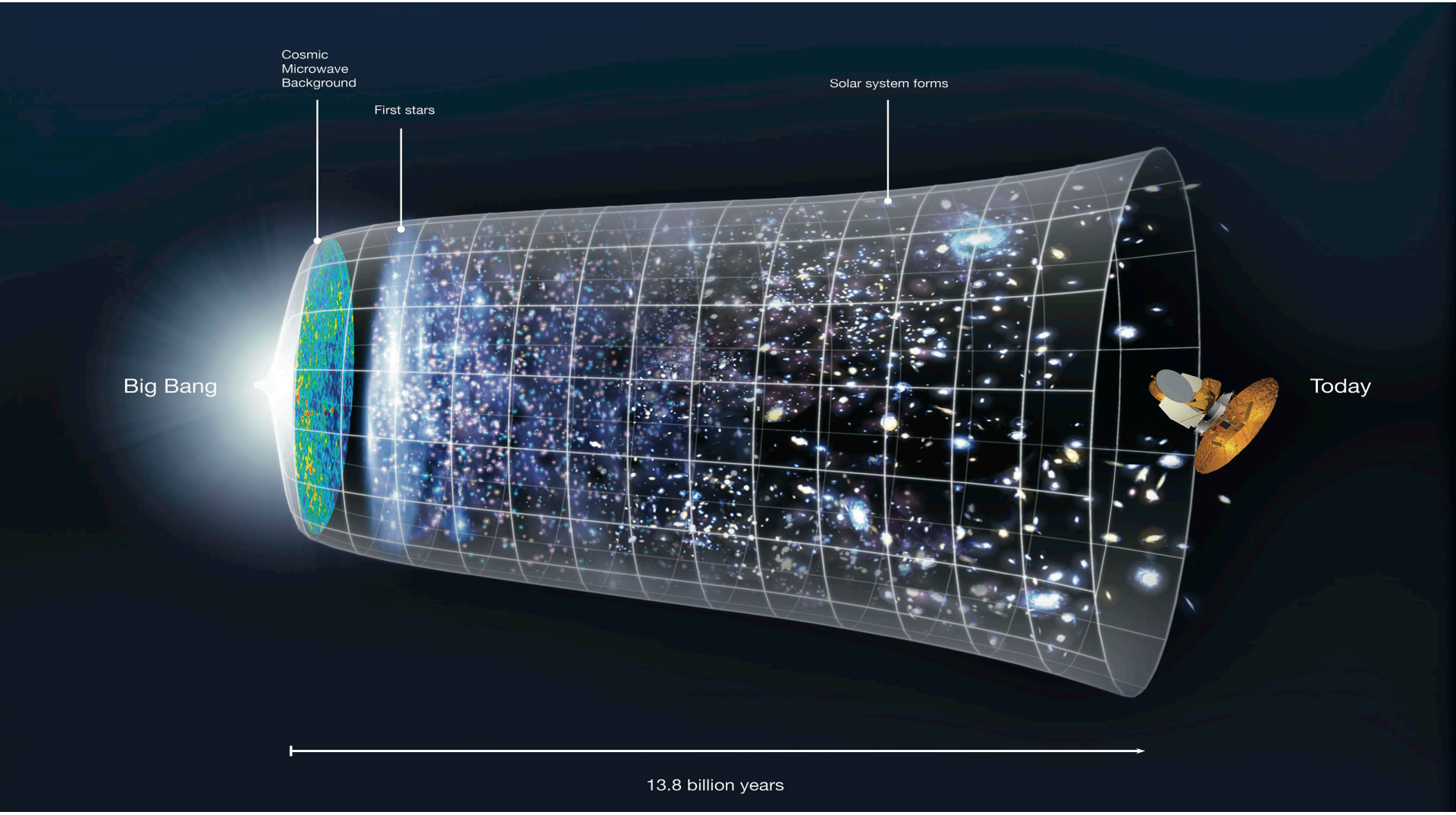
Romania	118

Associate Members in the Pre-stage to Membership

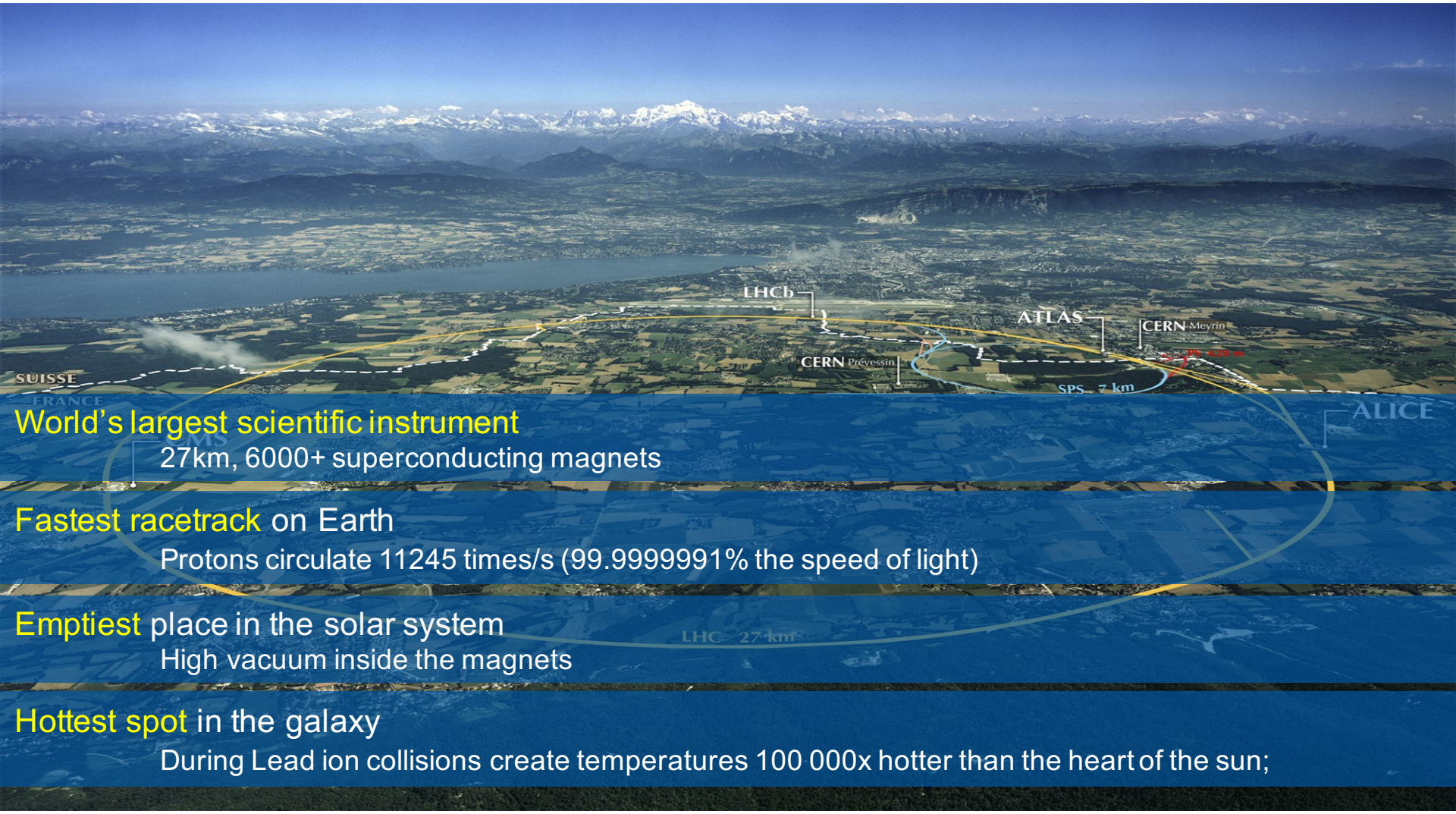
Serbia	41

Distribution of All CERN Users by Nationality on 14 January 2014

01. Fundamental Research – The Building Blocks of the Universe



01. The Large Hadron Collider – A Marvel of Technology



World's largest scientific instrument

27km, 6000+ superconducting magnets

Fastest racetrack on Earth

Protons circulate 11245 times/s (99.9999991% the speed of light)

Emptiest place in the solar system

High vacuum inside the magnets

Hottest spot in the galaxy

During Lead ion collisions create temperatures 100 000x hotter than the heart of the sun;

01. The Large Hadron Collider – An Extremely Complex Instrument

CERN Accelerator Complex is unique installation

Therefore, we have to face unique challenges

Control and Operations

Million of sensors, large number of control devices, front-end equipment, etc.

Many critical systems: Cryogenics, Vacuums, Machine Protection, etc.

01. The Experiments – ATLAS, CMS, ALICE and LHCb Detectors

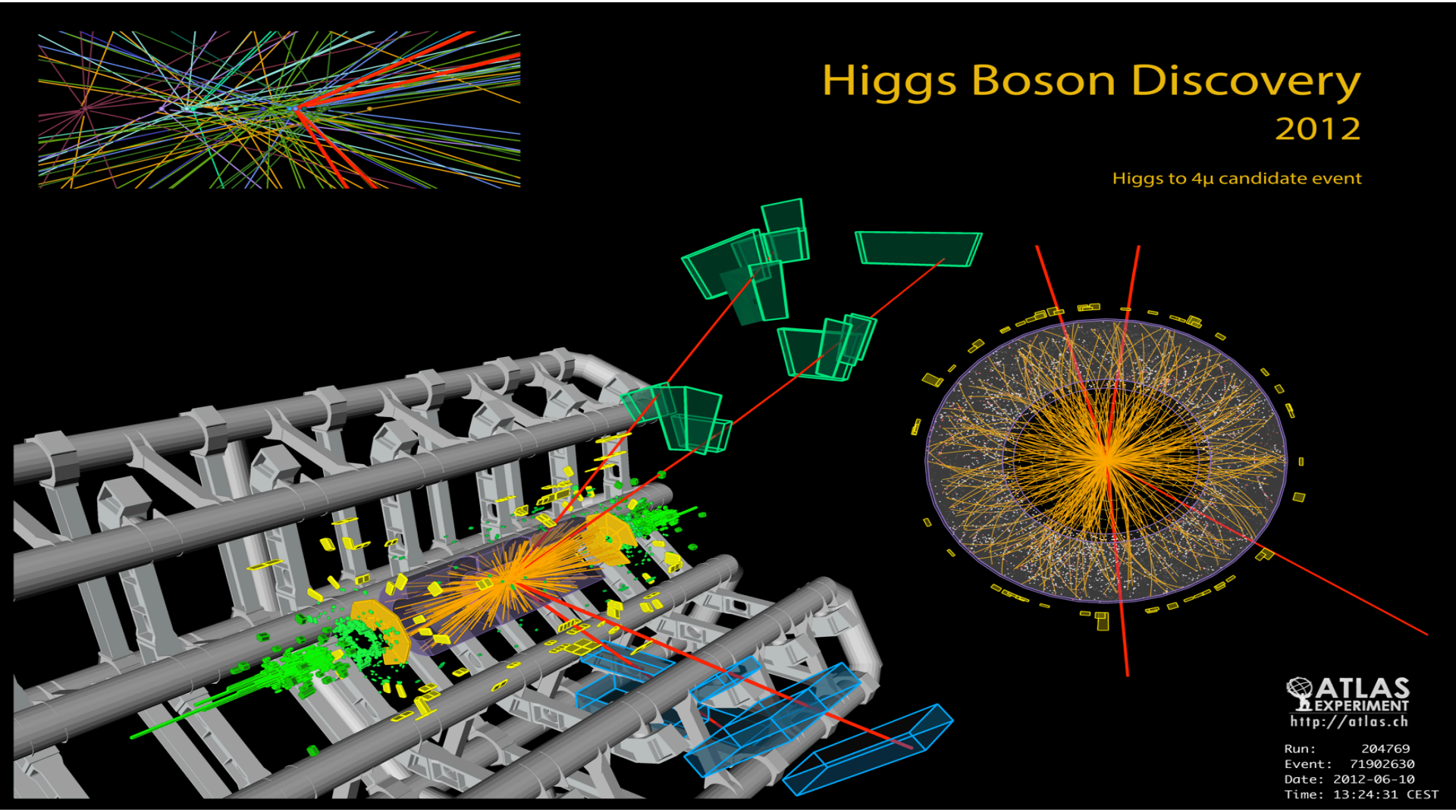
150 Million of sensor

Control and detection sensors

Massive 3D camera

Capturing 40+ million collisions per second

01. The Experiments - Millions of Collisions Per Second

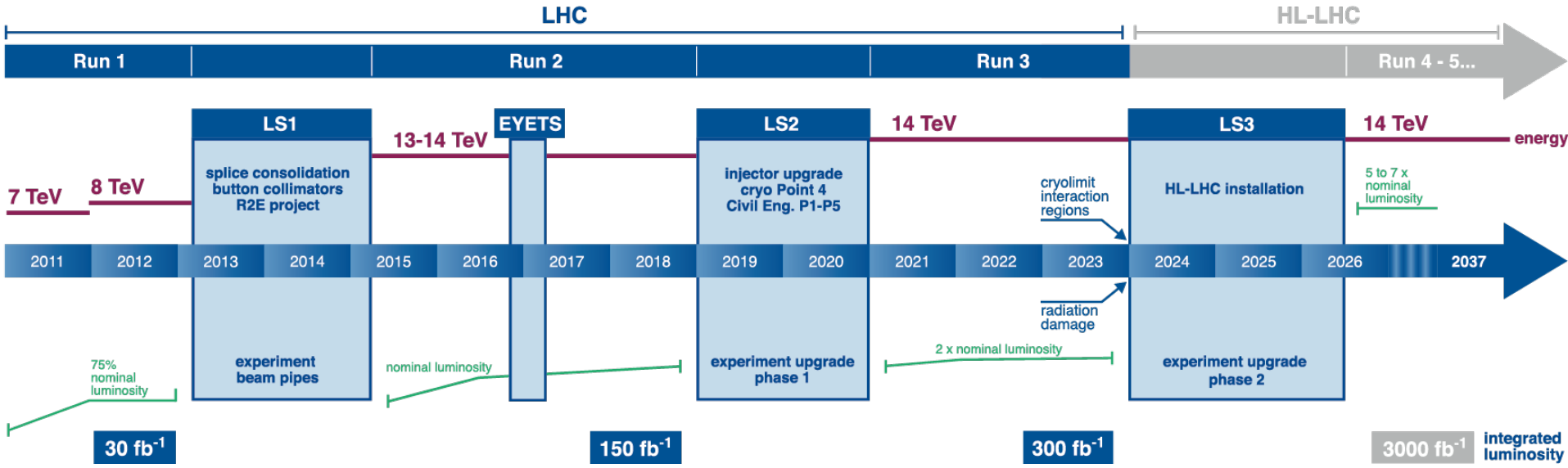


Agenda

1. CERN Introduction
2. **Why CERN cares about Big Data?**
3. What does Big Data looks like? What are the benefits and risks?
4. The data process: jumping into Big Data
5. Profiting from the data investments
6. Making the most of it: Advance and scalable analytics
7. Some conclusions from our experince.

02. What are the Challenges Ahead - Toward High Luminosity

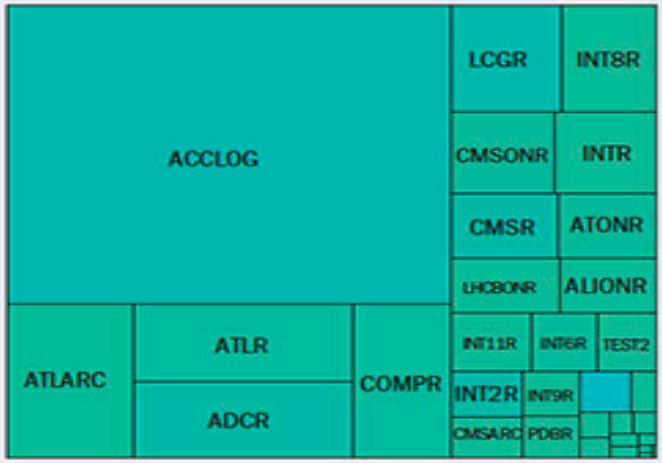
LHC / HL-LHC Plan

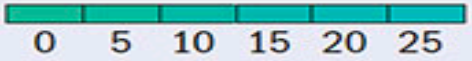


- The constant to push the machine to its limits
- Exiting future for scientists but a huge challenge for the engineers

02. CERN's Database Status – More Data and Faster

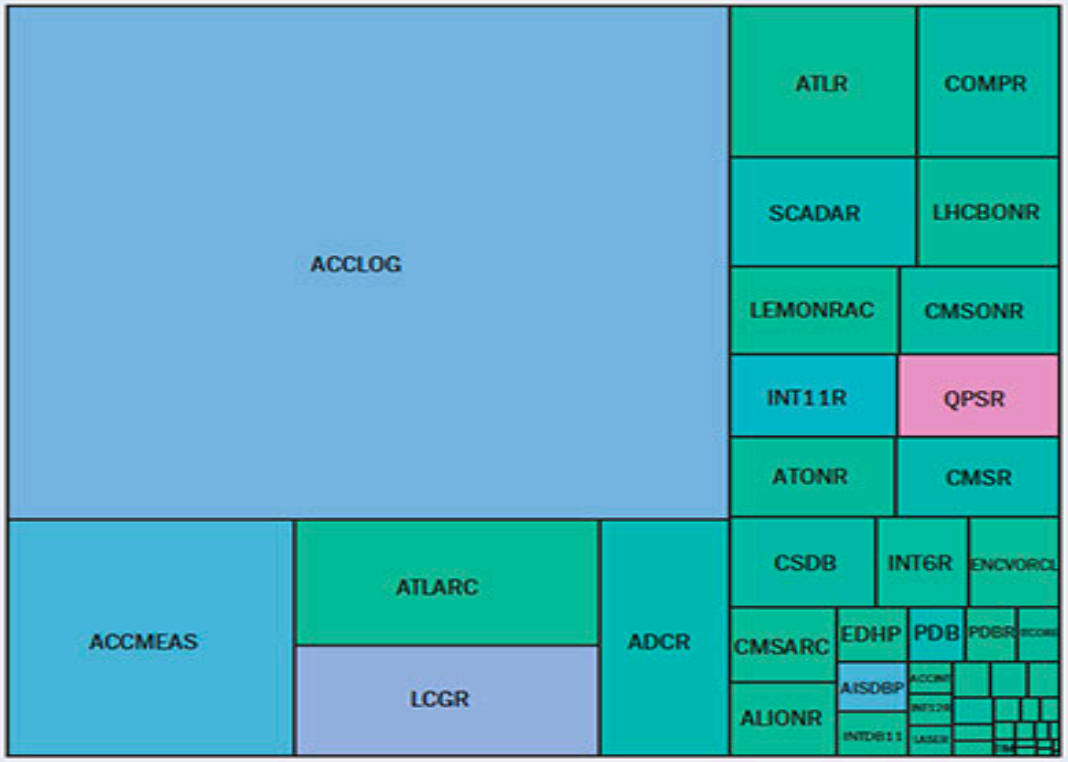
CERN databases October 2012
area is size (total = 299 TB)

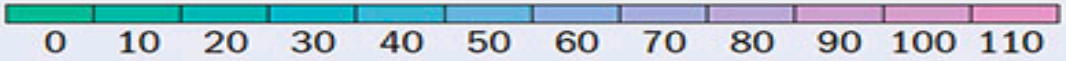




TB redo per month, sum = 120 TB

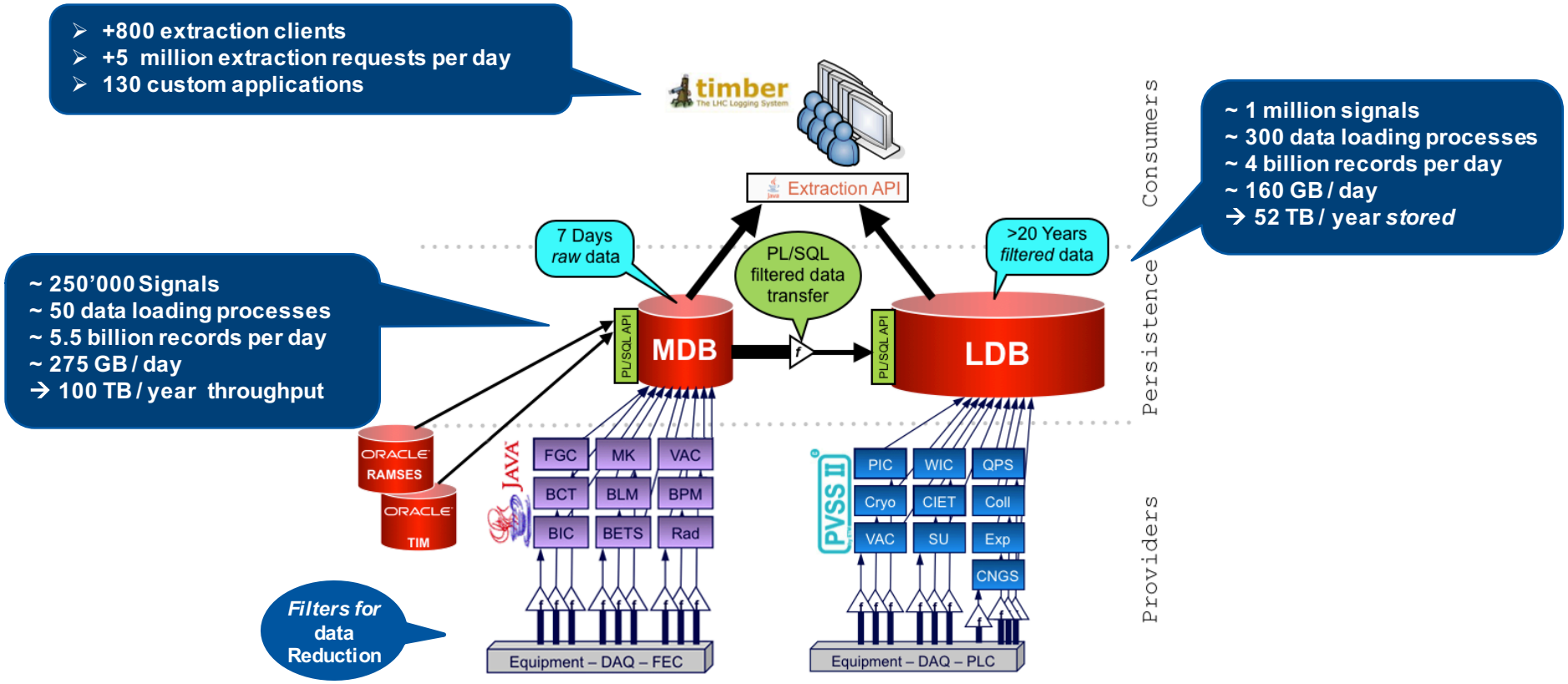
CERN databases December 2015
area is size (total = 750 TB), colour is redo activity



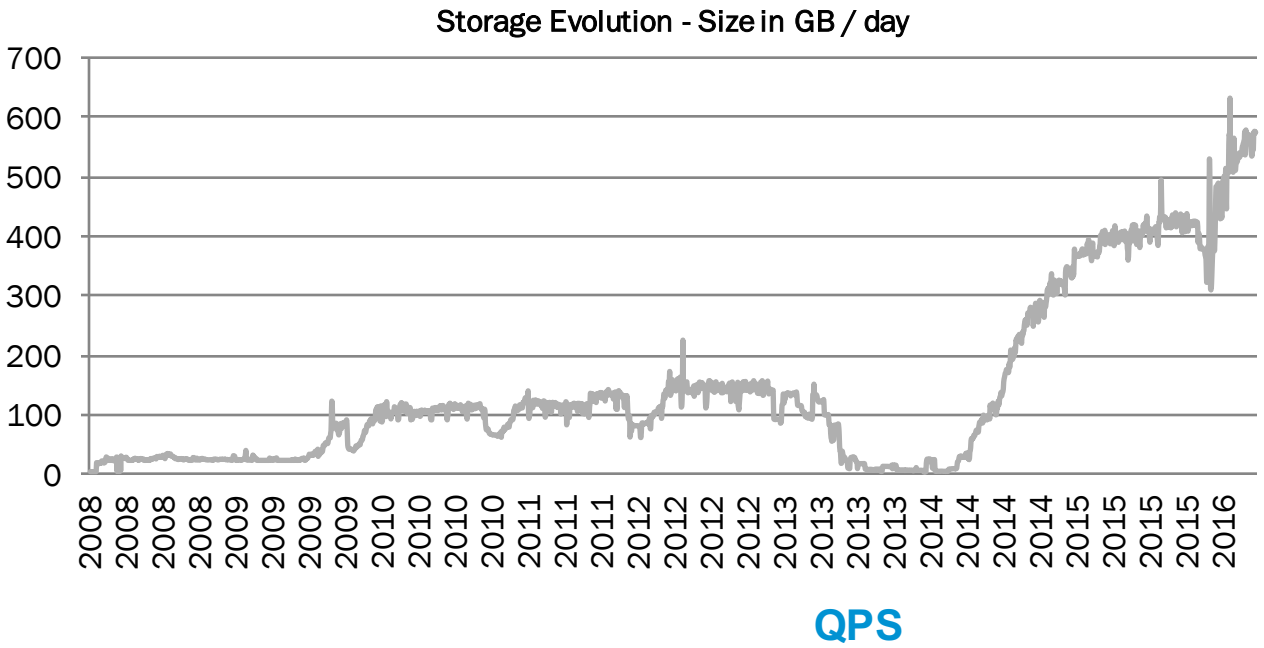


TB redo per month, sum = 494 TB

02. CERN's Accelerator Logging Service – Control IoT System



02. CERN's Accelerator Logging Service – Data Evolution



02. The Evolution Towards Big Data

What we do have…

- A cost-effective system
- The functionalities covers more than 90% of current requirements
- Data access with SQL, custom Java APIs and applications

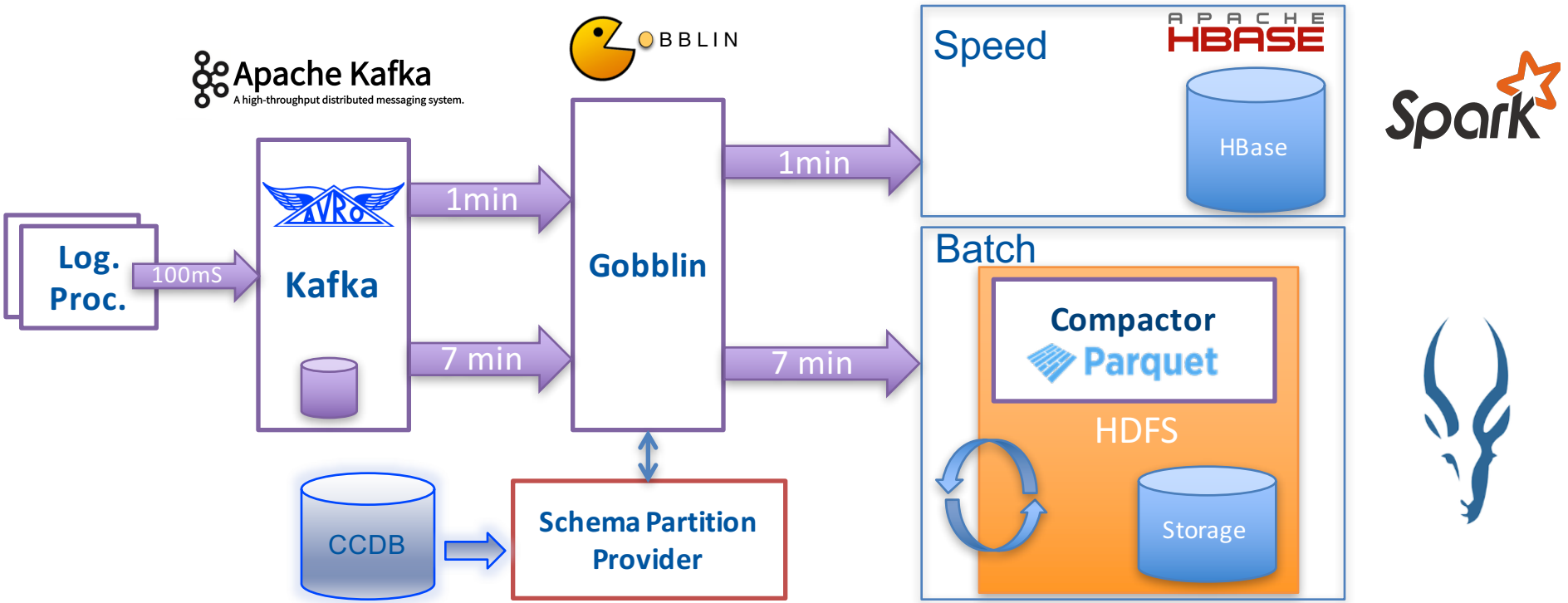
…then why to change?

- Better performance and unlock bigger datasets
- Leverage analytics capabilities: new data exploitation approaches
- More heterogeneous data access models

Agenda

1. CERN Introduction
2. Why CERN cares about Big Data?
3. What does Big Data looks like? What are the benefits and risks?
4. The data process: jumping into Big Data
5. Profiting from the data investments
6. Making the most of it: Advance and scalable analytics
7. Some conclusions from our experince.

03. New Technologies – New Challenges, Risks and Opportunities



03. New Technologies – New Challenges, Risks and Opportunities

The new architecture fulfils current and future requirements…

- Scalability
- Heterogeneity of clients:
  - Not limiting the accessing technologies
  - Different expertise
- More and better ways to exploit the data
  - Bigger datasets
  - Real-time or streaming analytics,
  - More complete and informative decisions
- Leverage Collaboration
  - Across different locations
  - Across academic and processional backgrounds

03. New Technologies – New Challenges, Risks and Opportunities

…but what are the costs and risks?

- Human and expertise
- Economical costs underestimation
- Overestimation of expectation
- Lack of maturity of the technologies to run critical services
- Evolution and changing pace of the technologies

03. New Technologies – New Challenges, Risks and Opportunities

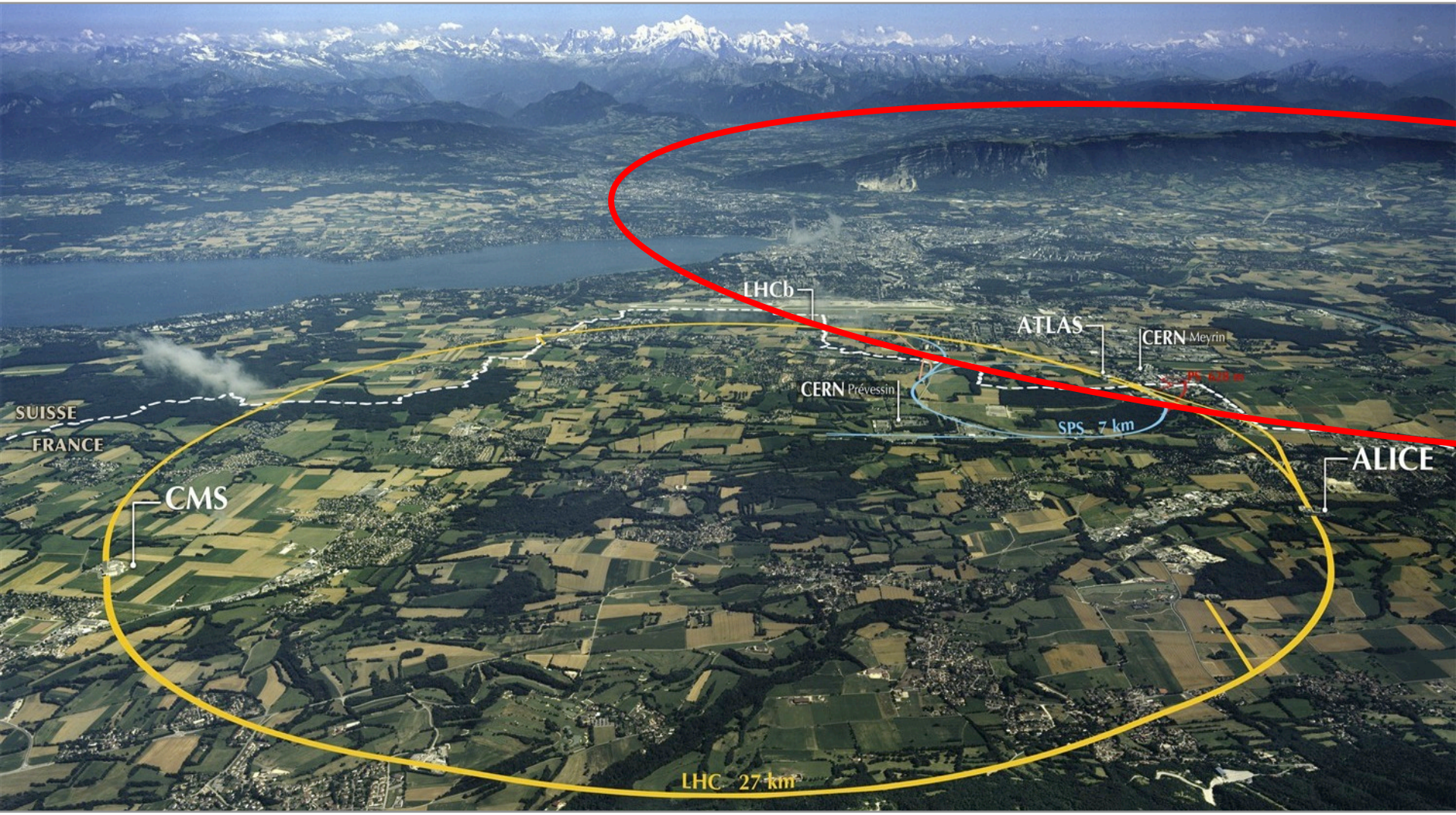
…how minimize them?

- Academic programs have quickly adapted to the expertise need
- Avoid attachments to technologies by decupling services
- Strategies to find the balance between productivity and exploration
- Same old problems will show up - use your expertise
- Cloud, Cloud and Cloud:
  - Fast deployments
  - Focus on solutions and not architecture
  - Minimize the costs

Agenda

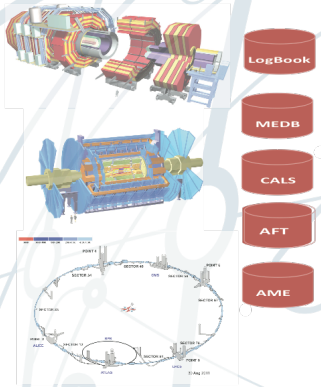
1. CERN Introduction
2. Why CERN cares about Big Data?
3. What does Big Data looks like? What are the benefits and risks?
4. **The data process: jumping into Big Data**
5. Profiting from the data investments
6. Making the most of it: Advance and scalable analytics
7. Some conclusions from our experince.

04. Future Circular Collider: Enhance Collaboration with Big Data



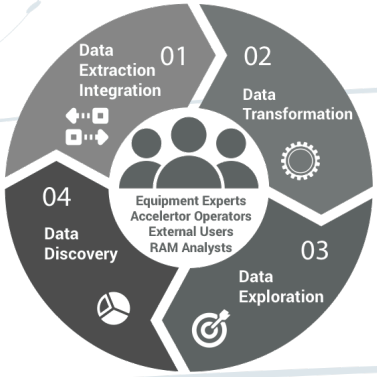
04. The Data Process: Making Data Accessible

Scenario



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

NoSQL
XML
JSON
Text
RDBMS

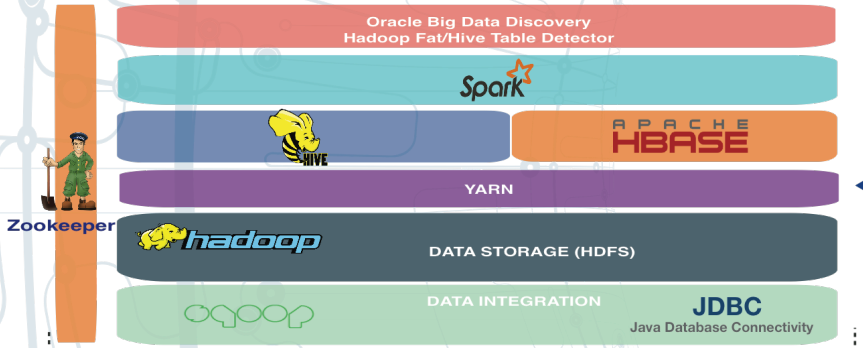


Challenges

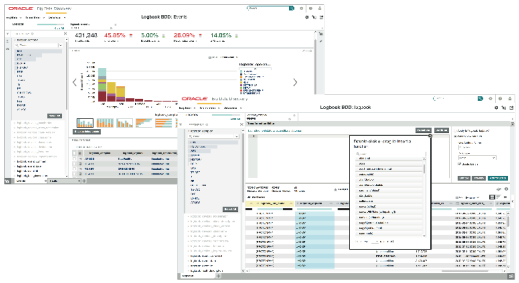
- Scale-out architecture
- Based on cloud
- Advance Analytics (Text Analytics, Correlation)
- User-driven ETL
- Flexible Data Analytics and Discovery
- Collaborative environment

Architecture

ORACLE cloudera



Oracle DGRAPH
In-Memory Columnar Database



Oracle Big Data Discovery Studio

Data persistency and ETL

Data Exploration

Data Discovery and Analytics

04. The Process: Offering the Right Tools – Data to Value

Data Analysis, visualization and discovery involves several steps:

- Data integration, transformation, exploration, and discovery
- Agile data governance

Reliability Availability & Maintainability for CERN's Accelerator Complex

Import and export specific data in various formats and save them directly to Hadoop

Powerful editor for data transformation, data cleaning, attribute enrichment and feature engineering

Multiple built-in analytics components to create discovery applications and dashboards using one or several datasets.

Flexible environment allow users to create and customize their applications to dynamically filter and dig into the data to get insight

Agenda

1. CERN Introduction
2. Why CERN cares about Big Data?
3. What does Big Data looks like? What are the benefits and risks?
4. The data process: jumping into Big Data
5. Profiting from the data investments
6. Making the most of it: Advance and scalable analytics
7. Some conclusions from our experince.

05. The Data Process - Data to Value how to…

The goal…

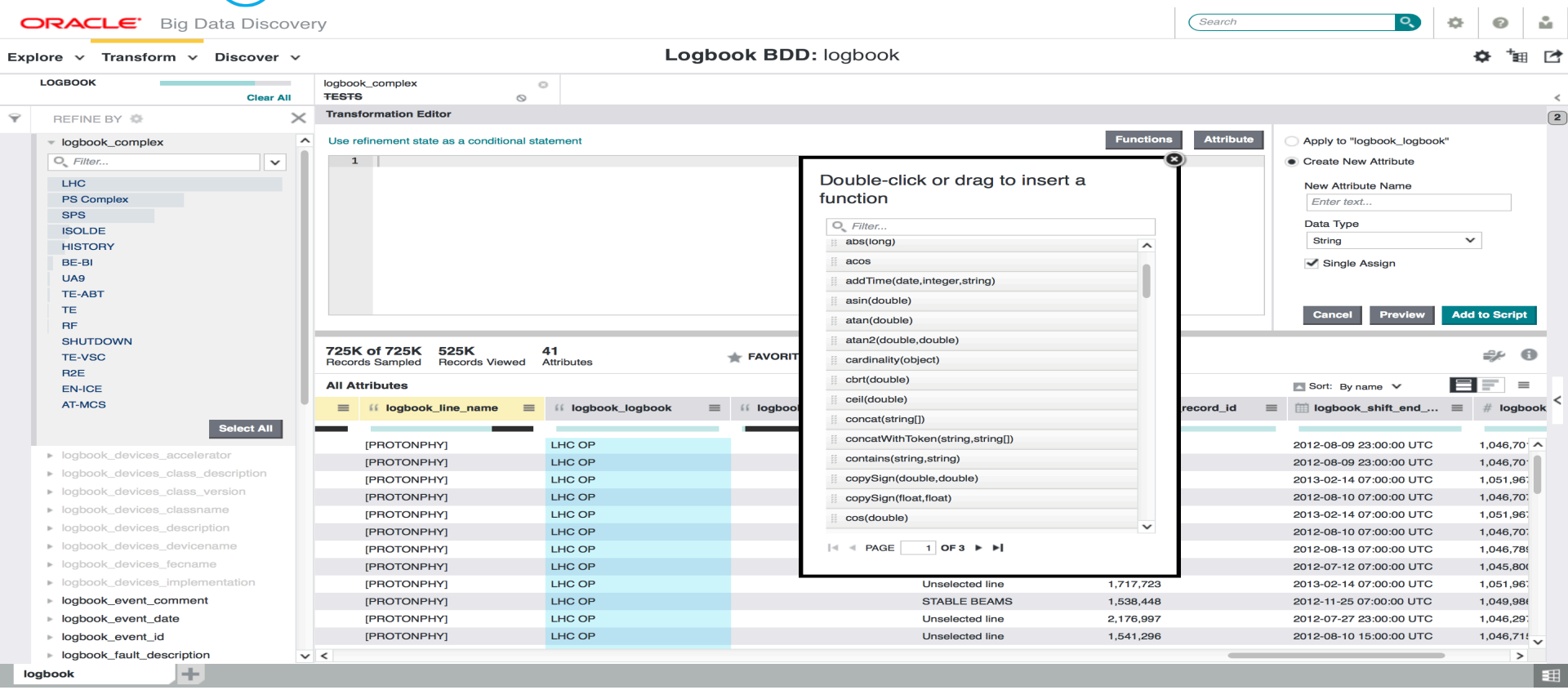
- Decentralize the data process
- Push the responsibility of the process to the people who better knows it

…but is that possible?

- Use the right tools – flexible, intuitive and almost-zero learning curve
- Hide the technology complexity
- Help to focus on the data that matters
- Allow a collaborative environment

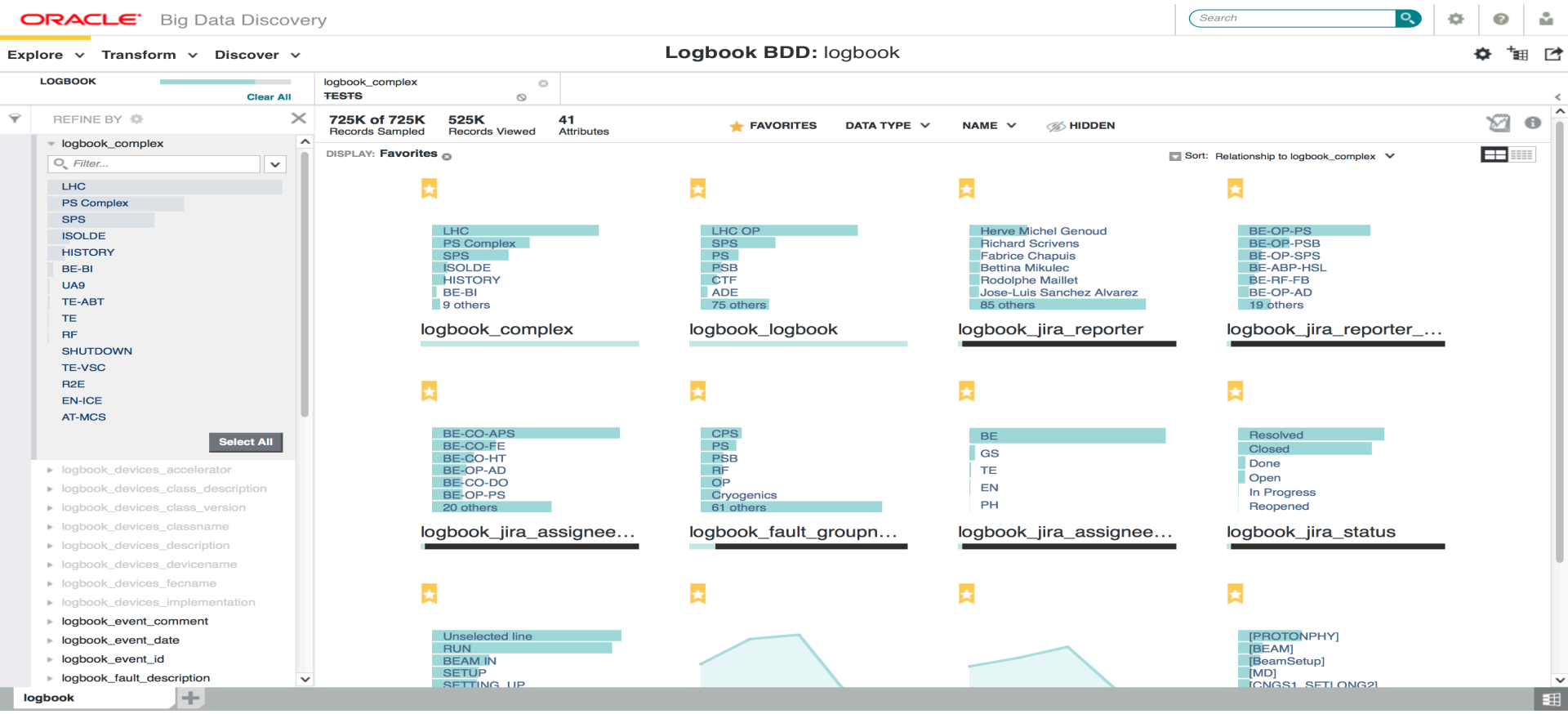
05. The Data Process - Hiding Technology Complexity

- Unlock the exploitation of data from different perspectives
- Avoid technological barrier
- Add an abstraction layer that minimize the risks



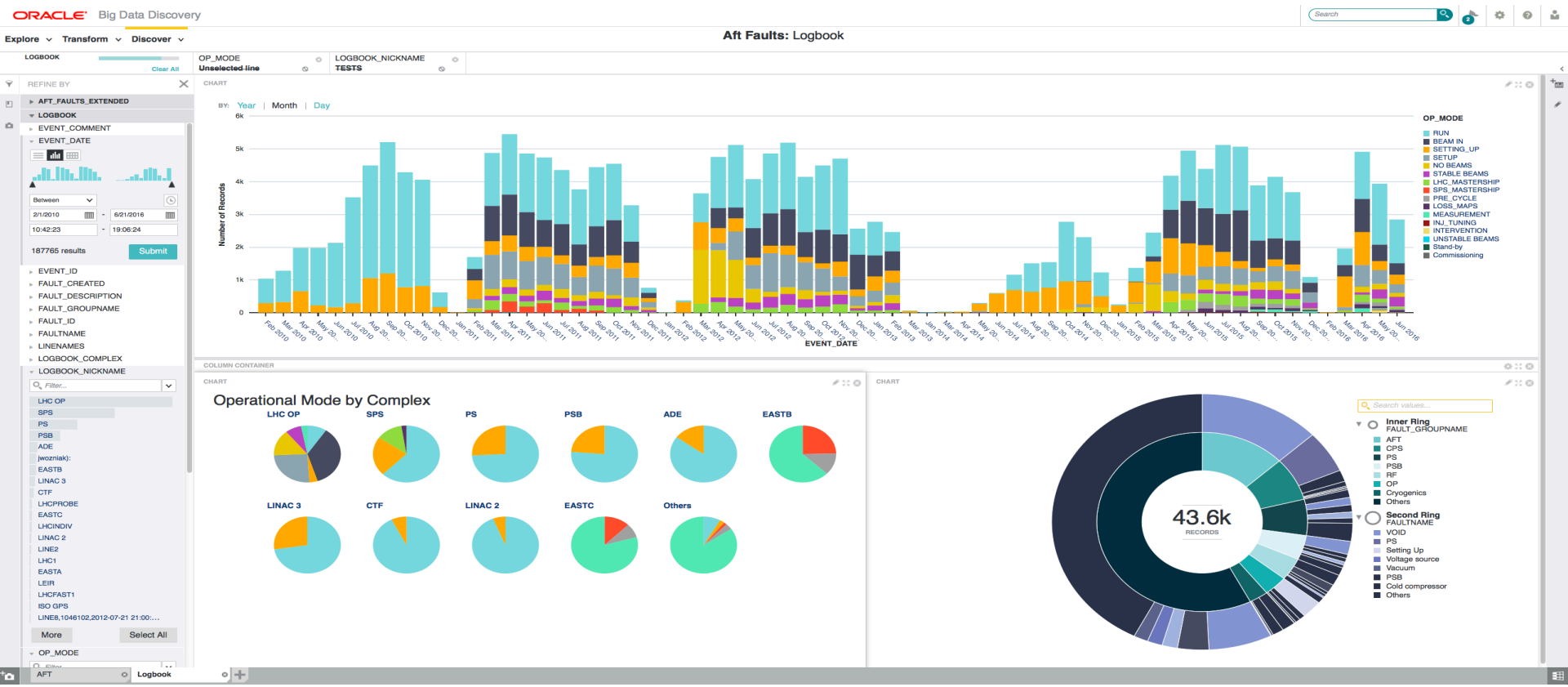
05. The Data Process - Fishing into the (Data) Lake

- Help to focus on the data that matters
- What do I really have available?



05. The Data Process - Sharing the Findings

- Use the right tools – flexible, intuitive and almost-zero learning curve
- Push the responsibility to explore the data to the people who better knows it



05. The Data Process - Where We Are and Where Are We Heading

So we already have achieved…

- Making the data available and integration of data sources
- The data is explored by the people who better knows it
- We can easily share findings and explorations

…but all that is part of the advance analytics process

- It helps to focus on the data that really matters
- Allow more effective advance analytics prototypes

Row Data -> Integration ->Transformation -> Focus -> Create Prototype

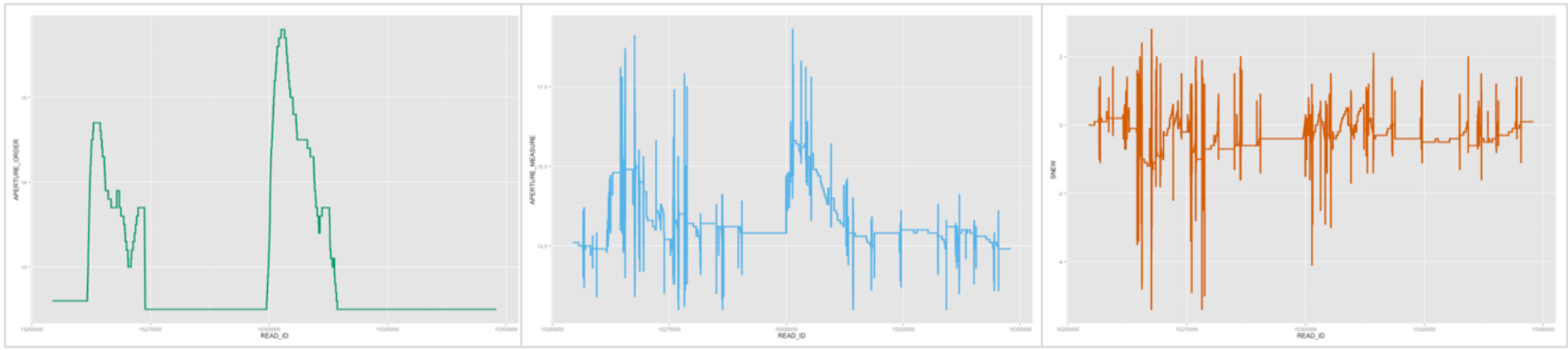
Agenda

1. CERN Introduction
2. Why CERN cares about Big Data?
3. What does Big Data looks like? What are the benefits and risks?
4. The data process: jumping into Big Data
5. Profiting from the data investments
6. **Making the most of it: Advance and scalable analytics**
7. Some conclusions from our experince.

06. Making the most of it: Advance Analytics

Lets apply predictive maintenance to the LHC control system..

Aperture order | Aperture measured | **Difference = measured - order**

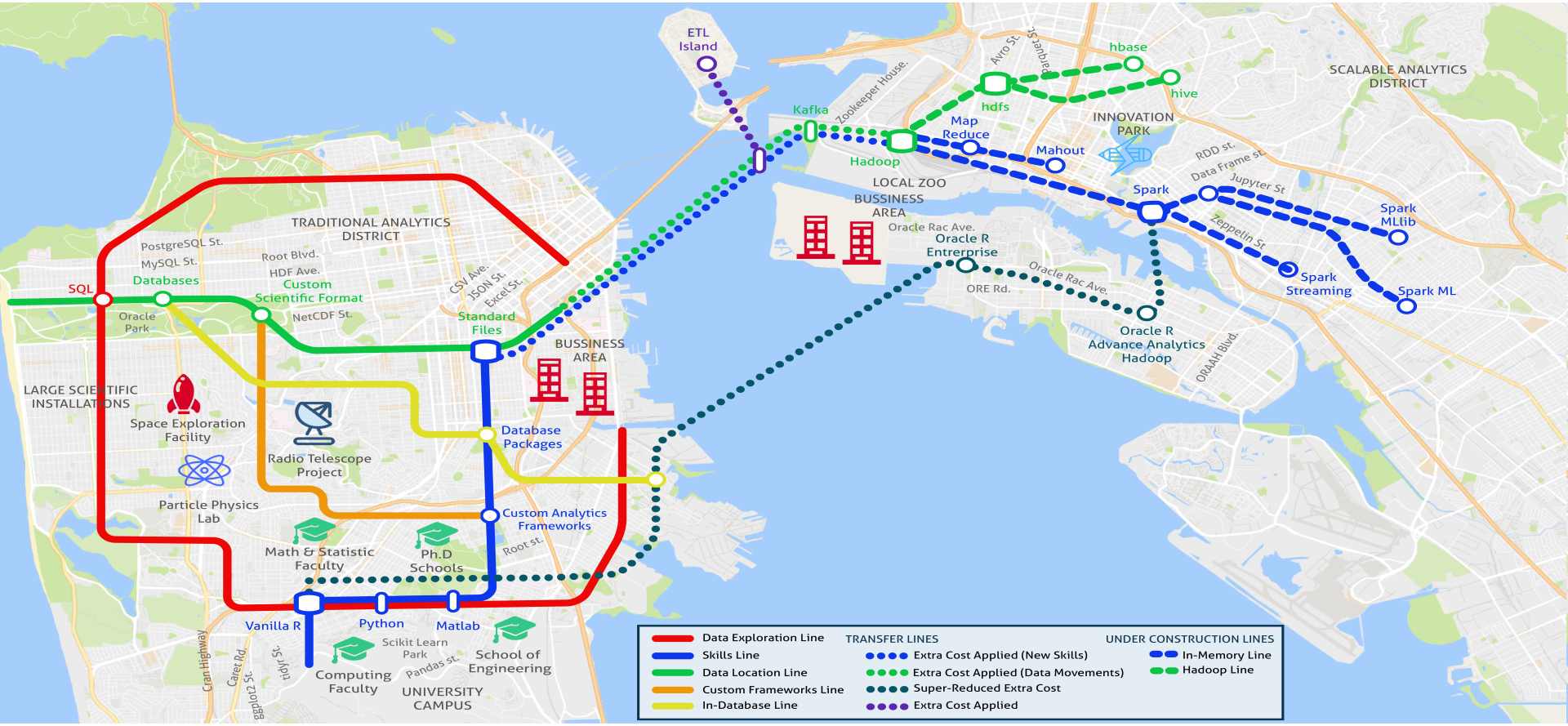


Row Data -> Integration ->Transformation -> Focus -> Create Prototype

06. Making the most of it: Scalable Advance Analytics

…nice prototype but now I need to scale

Need for a new set of skills and therefore profiles

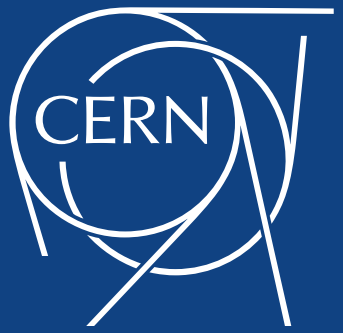


Agenda

1. CERN Introduction
2. Why CERN cares about Big Data?
3. What does Big Data looks like? What are the benefits and risks?
4. The data process: jumping into Big Data
5. Profiting from the data investments
6. Making the most of it: Advance and scalable analytics
7. Some conclusions from our experince.

05. Some Conclusions

- Embark into Big Data is a risk but the risk of not doing it is much higher
- Unlocks new ways to exploit your investment on Data
  - Overcomes technical limitations
  - Allows more heterogeneous data access
- But making the data is available is just half of the way
  - Profit from data is a full process
  - Self-service tools enhance collaboration
  - Reduce IT intervention improve productivity
  - Scale your advance analytics
- Cloud is essential
  - Fast and flexible deployments
  - Focus on solutions and not architecture
  - Minimize the risks and costs
- And please do not underestimate costs or overestimate expectations


CERN