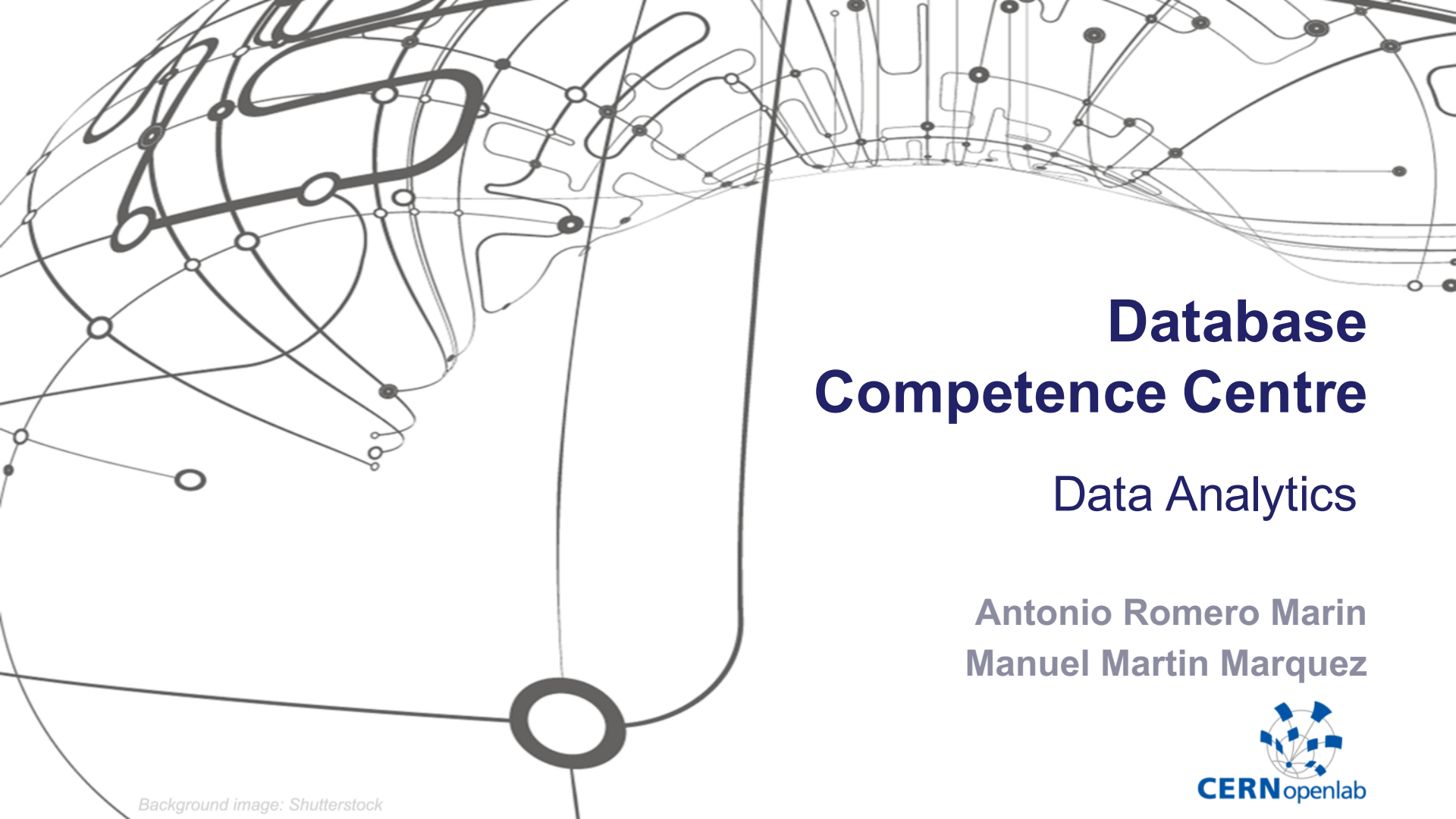


Database Competence Centre

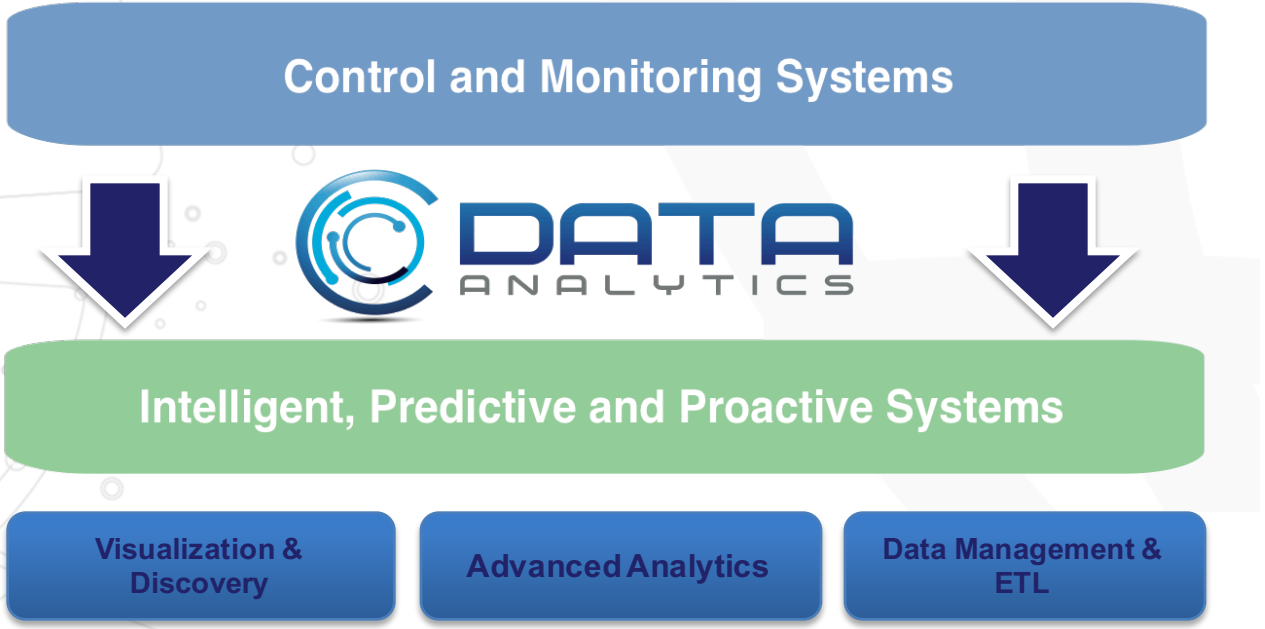
Data Analytics

Antonio Romero Marin
Manuel Martin Marquez

CERN openlab

Background image: Shutterstock

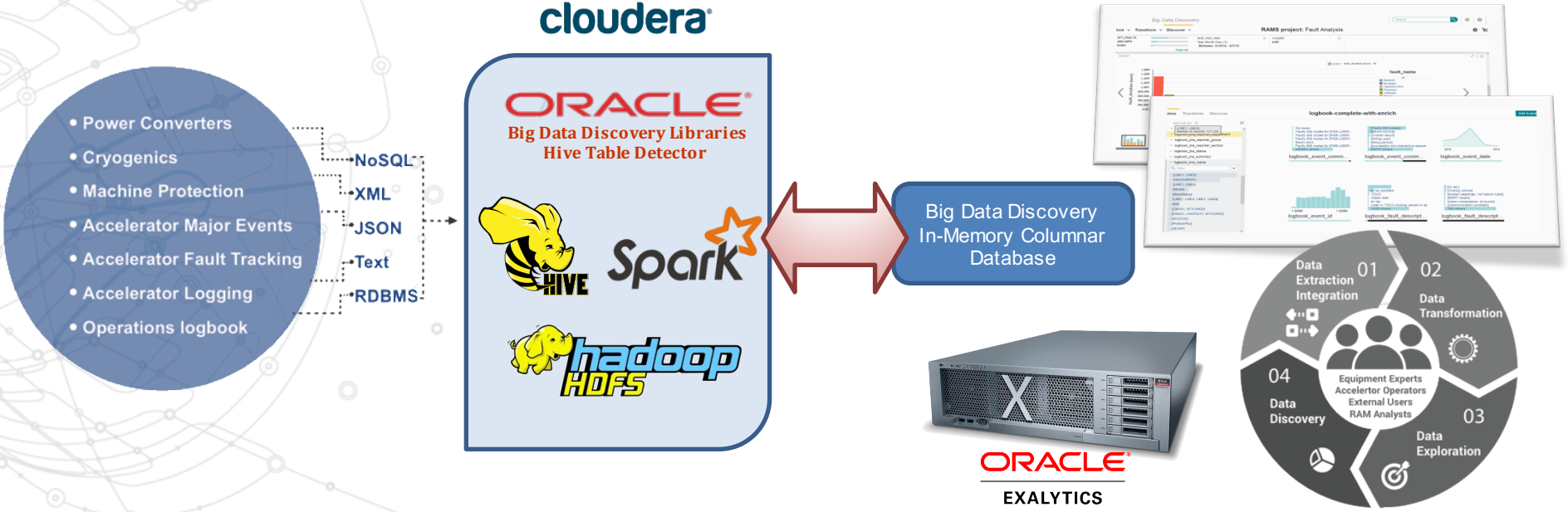
Data Analytics Scenario



Visualization and Discovery

- **Goals**
  - Visualize any type of data
  - Easy to use and self-service
  - Collaborative environment
- **Recent Activities**
  - From Endeca Information Discovery to Big Data Discovery (BDD)
  - Evaluation of BDD 1.0 (March 15)
  - Evaluation of BDD 1.1 (October 15)
  - Installation on Exalytics machine (October 15)
  - PoC – 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

Discovery platform overview



Background image: Shutterstock

Oracle Big Data Discovery Benefits

- **Self-Service Data Transformation, Exploration & Discovery**
  - Develop discovery applications and dashboards fast and easy
  - Interactive and flexible data visualization and exploration
  - Apply transformations and enrich the datasets
  - Not depend on IT to modify applications or datasets
- **Integrate any type of data sources and add relations**
  - Get a full picture of data originally spread in multiple data sources
  - Close integration with Hadoop and RDBMS
- **Collaborative environment**
  - Bookmark findings and share them with the rest of the team

Advanced Analytics

- **Goals**
  - Predictive maintenance and system optimization
  - Batch and real time analytics
- **Recent Activities**
  - Data Analytics with Oracle R Enterprise
    - Allows in-database data analysis
      - Providing scalability and performance, eliminating memory constraints
      - Most of our data sources are in Oracle RDBMS
    - Use cases
      - Predictive maintenance and anomaly detection on sensor data
  - Evaluation of Apache Spark for Data Analytics at CERN
    - Collaboration with CMS to optimize CMS data replication (data popularity)
    - openlab summer student project: Siddha Ganju

Data Management & ETL

- **Goals**
  - Integration with existing data repositories
  - ETL at scale
  - Facilitate ETL process for IT and users
- **Recent Activities**
  - Framework to ingest data from RDBMS to Hadoop
    - Built around Apache Sqoop
    - Address CERN requirements
    - Summer student project by Aniridha Bose & Daniel Stein
    - Collaboration between IT-DB and CMS
  - Self-service ETL with Oracle Big Data Discovery

Next Steps

- **Discovery and Visualization**
  - Oracle Big Data Discovery
    - More advanced visualizations
    - Scalability
    - Machine learning integration
    - Data streaming for real time visualization
- **Advanced Analytics**
  - Oracle R Enterprise
    - Integration with Spark and Hadoop
  - Oracle Complex Event Processing and Stream Explorer
    - Find patterns and anomalies in real time from stream sensors data
- **Data Management and ETL**
  - Continue development of ELT framework
    - Improve automatization
    - Address more complex use cases

Outreach

- 13 external presentations and 1 poster
- Article published in Swiss Analytics magazine
- Poster at openlab Open Day
- **Oracle OpenWorld 2015**
  - EMEA Big Data breakfast presentation
  - Oracle Big Data Discovery panel session
  - Big Data and Data Warehousing panel session
  - Executive Summit: Big Data panel session
  - Video Interview on Big Data Discovery
- More details: **http://openlab.web.cern.ch/publications**

www.cern.ch/openlab