ECB-PUBLIC
FINAL

EUROPEAN CENTRAL BANK

EUROSYSTEM

Markus Trzeciok
Data Analytics and Domain Services
DG Information Systems

Juan Alberto Sánchez
Statistical Applications and Tools
DG Statistics

Big data and Machine Learning initiatives at the ECB

*** The views expressed here are those of the presenters and do not necessarily reflect those of the ECB.**

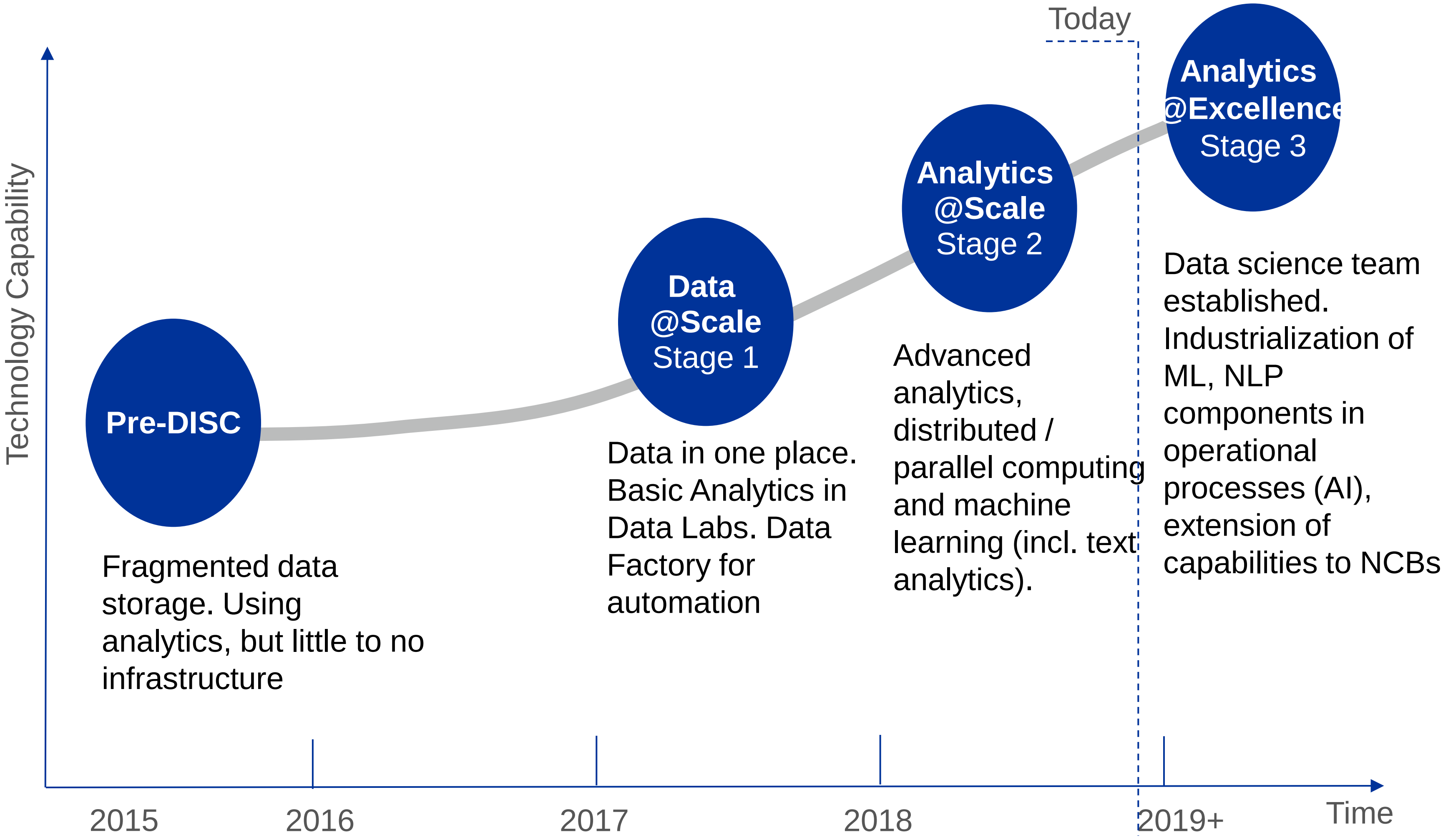
Bank of Italy and BIS Workshop on “Computing Platforms for Big Data and Machine Learning”

Rome, 15th January 2019

Overview

1. **Becoming a cognitive organization**
2. Architecture and IT Services
3. Data Science Services & Activities
4. Strategy for data on boarding and Integration
5. First experiences working with Big Data Platform: EMIR & SUBA
6. Machine Learning in DG-Statistics

The journey to become a cognitive organization



Pre-DISC

Fragmented data storage. Using analytics, but little to no infrastructure

Data @Scale
Stage 1

Data in one place. Basic Analytics in Data Labs. Data Factory for automation

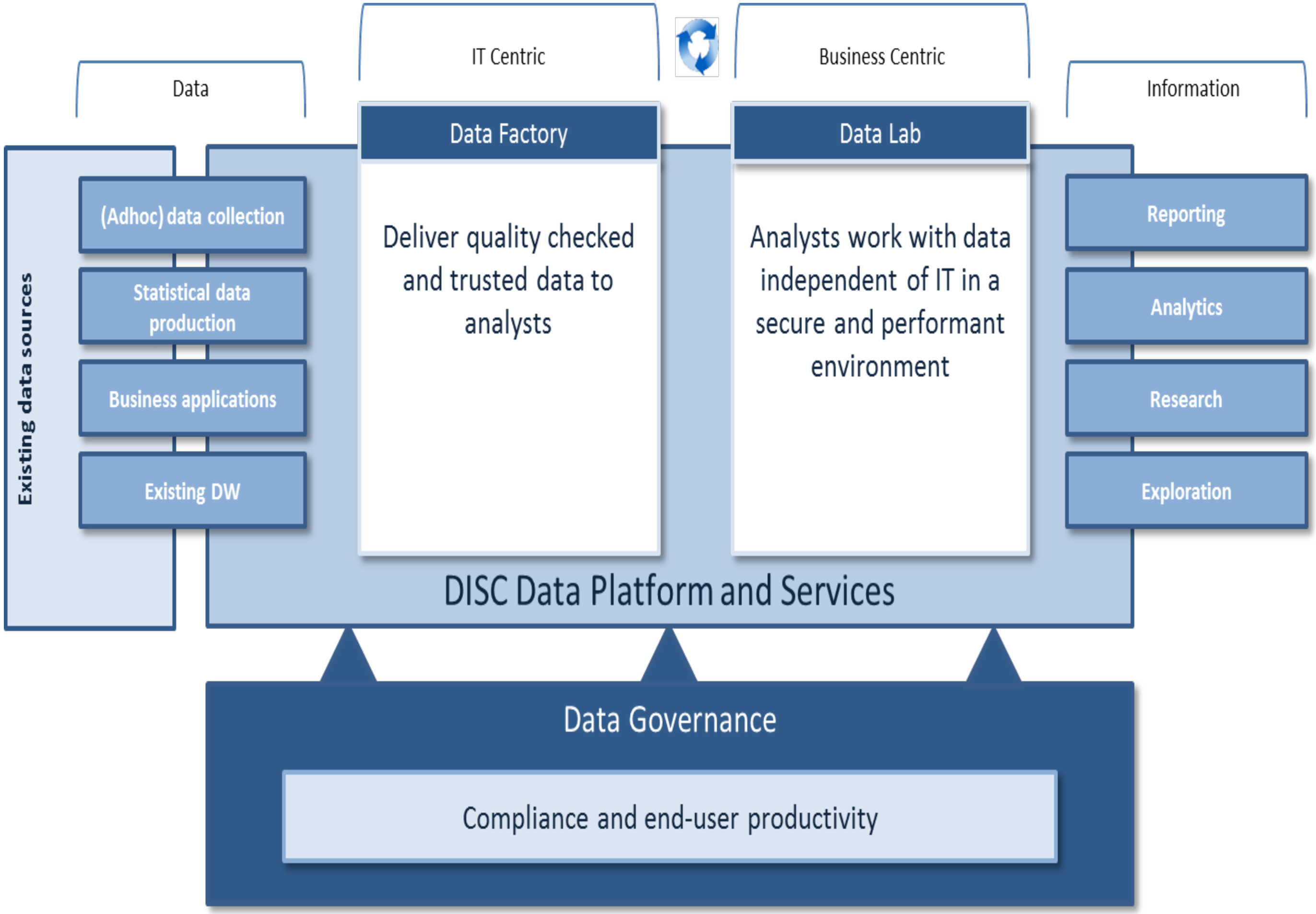
Analytics @Scale
Stage 2

Advanced analytics, distributed / parallel computing and machine learning (incl. text analytics).

Analytics @Excellence
Stage 3

Data science team established. Industrialization of ML, NLP components in operational processes (AI), extension of capabilities to NCBs

Conceptual Architecture


IT Centric
Business Centric
Data
Information
Data Factory
Deliver quality checked and trusted data to analysts
Data Lab
Analysts work with data independent of IT in a secure and performant environment
Existing data sources
(Adhoc) data collection
Statistical data production
Business applications
Existing DW
Reporting
Analytics
Research
Exploration
DISC Data Platform and Services
Data Governance
Compliance and end-user productivity

Self-Service Toolbox

Analytical Tools (laptop)

Source Code Management

Data Lab

Data Lab is like an empty database. Experts can load data files and create database tables and views without involvement of IT. Analytical tools can connect to Data Labs for programming and visualisation.

Data Lab Governance established

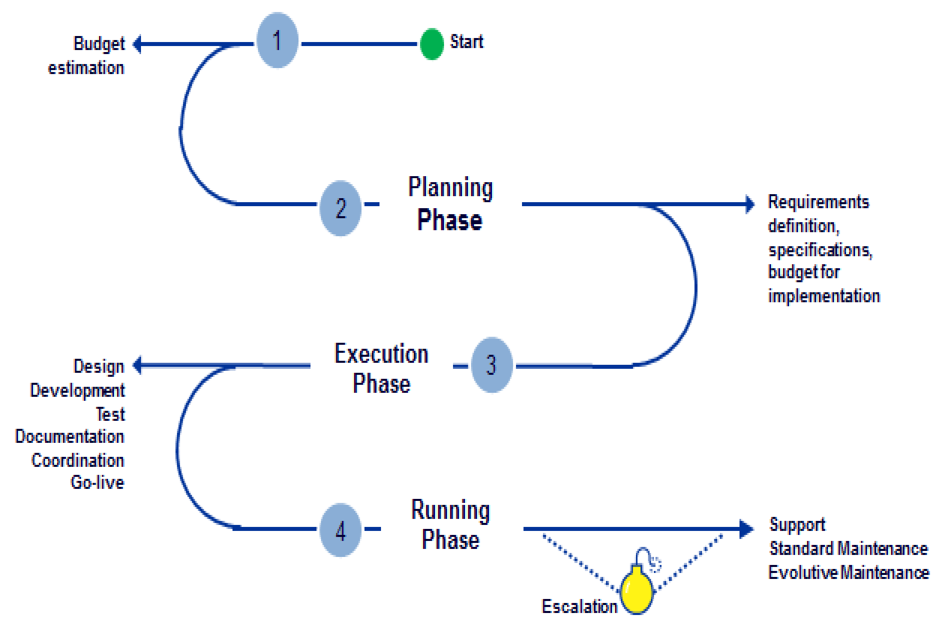
Data Science Workbench

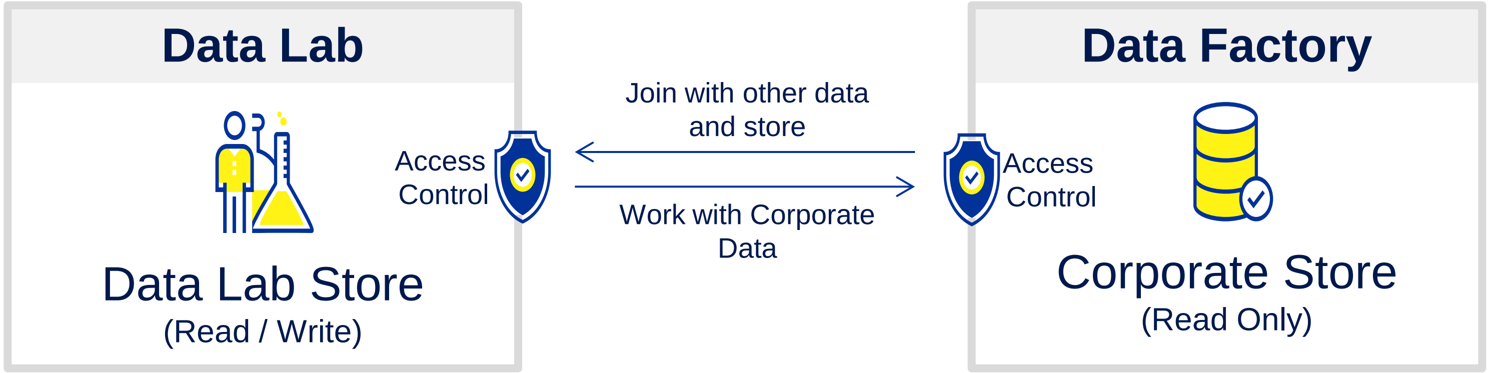
It is a development and runtime environment based on a computer cluster for python, R and Scala. Access to data in Data Lab is available as well as DISC Corporate Store. Native integration with Bitbucket and scheduler to semi-automate workloads and processes.

IT Centric Service

Data Factory

Is a service to deliver datasets, data products, reports and dashboards. Data Factory services are used by projects / activities to on-board their datasets and to develop dashboards and reports with Tableau and BOSS.





Data Lab

Access Control

Data Lab Store

(Read / Write)

Join with other data and store

Work with Corporate Data

Data Factory

Access Control

Corporate Store

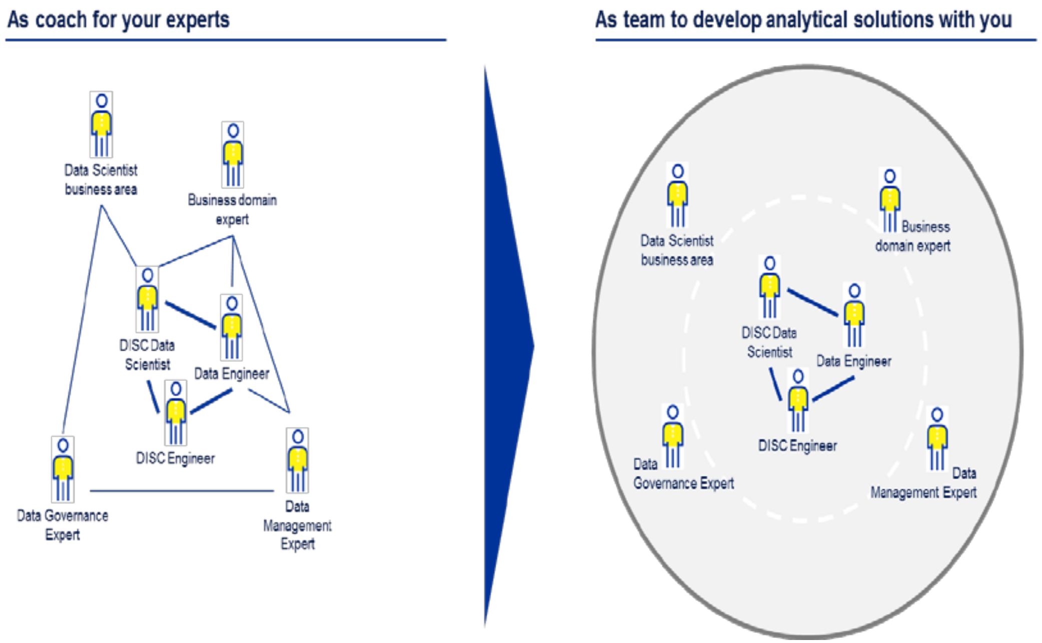
(Read Only)

Data Science

Data Science Nucleus

Production and analysis of data are at the heart of our decision making processes. The ECB is a house of data scientists.

The nature of data and technology is fast evolving. DISC provides services to master the rich and diverse toolbox available for data scientists.



As coach for your experts

As team to develop analytical solutions with you

Ad hoc support
Business experts develop their analytical solution on their own. Data science nucleus is available for ad hoc engineering and conceptual questions.

Structured support
Business experts develop their analytical solution on their own. Data science nucleus is available for code reviews, pair programming, coaching.

Solution development
DISC Data Science Nucleus develops the analytical solution in close collaboration with business experts.

Data Science Activities

Inflation nowcasting

Provide near real-time information (through web-scraping of online stores) on special factors inducing volatility to the inflation forecast (instead of explaining such deviations retrospectively); and second, conduct policy-relevant research. ML/NLP used for product classification according to COIPCO, DISC Cloud environment.

Mini Journey

D-BN started to collect sensor information from a sub-set of banknote machines. This information shall be used for various use-cases. For example, prediction of banknotes production, predict deterioration of banknote fitness, circulation of banknotes etc.

Legal opinions & SSM FAQ

Apply NLP and ML techniques integrated with SOLR for topic classification of legal opinions and SSM FAQ content. Aim is to improve search ability of content (a) to facilitate the consistent drafting of legal opinions by legal experts and (b) have faster access to relevant SSM FAQ content.

HR Analytics

HR is building an Analytics function which – in the first place – focusses on deriving value from existing data by providing intuitive report and dashboards. In the next step the aim is to apply advanced techniques (AI) and integrate with operational processes for staff mobility recommendations, applicant prediction, modelling demographical development.

Overview

1 Becoming a cognitive organization

2 Architecture and IT Services

3 Data Science Services & Activities

4 Strategy for data on boarding and Integration

5 First experiences working with Big Data Platform: EMIR & SUBA

6 Machine Learning in DG-Statistics

Strategy for data on-boarding and integration

Objective:

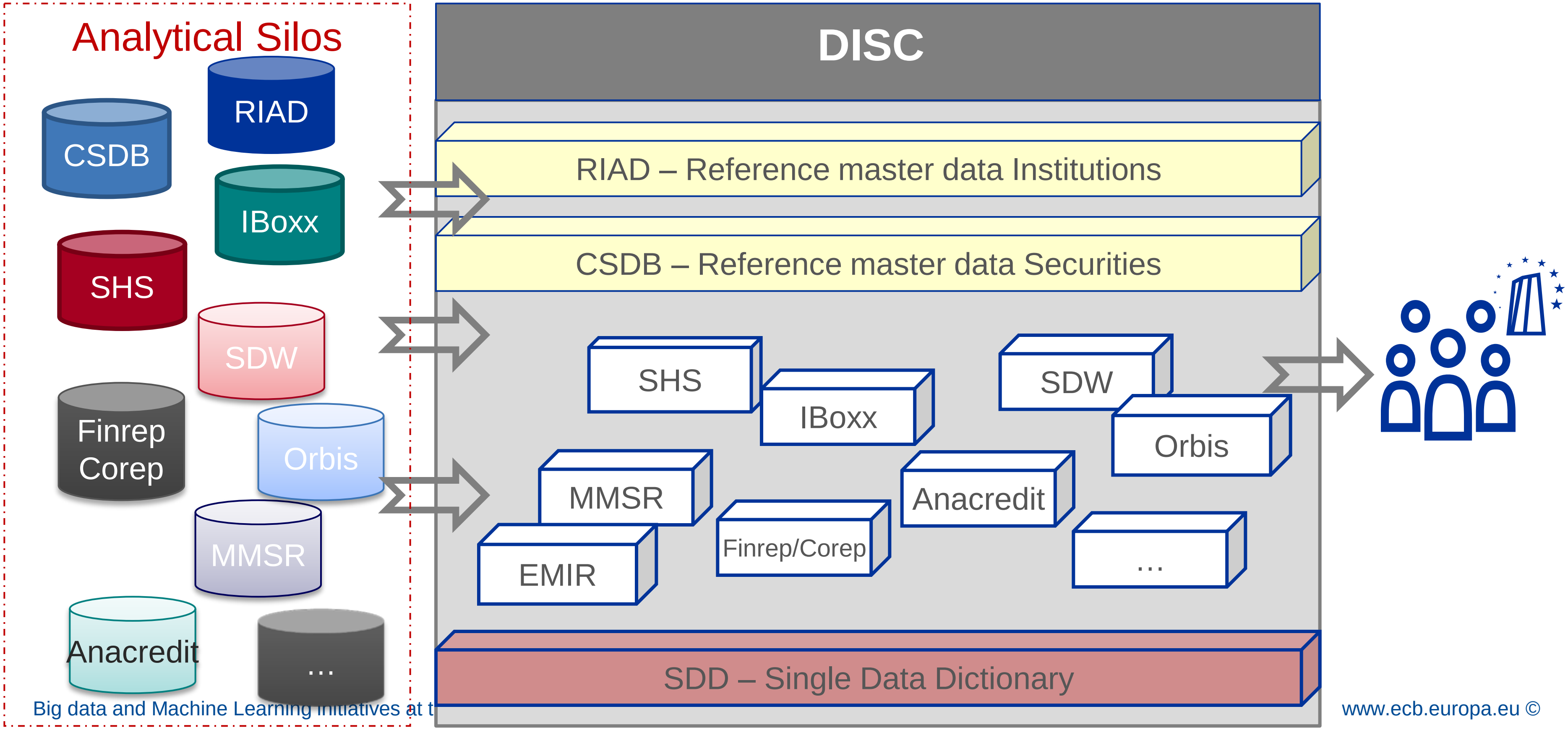
Facilitate analysis through the provision of integrated datasets in a common and powerful big data platform

- Big data technologies are an effective tool to support analysis based on granular data, for economics and financial stability
- Integrate datasets to facilitate analysis – Unified view
  - Dictionaries
  - Master Data
  - Empowering users and analytical capabilities
- Enable Advanced Analytics
  - Empower users
  - Machine Learning techniques

Strategy for data on-boarding and integration

ECB-PUBLIC
FINAL

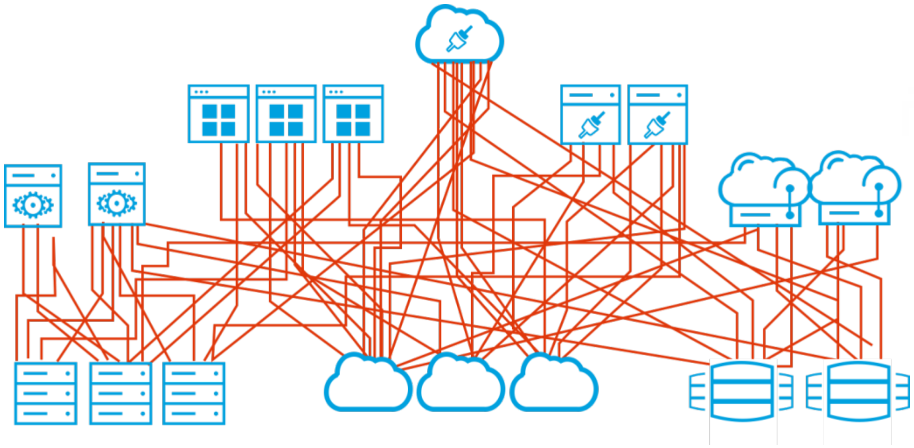
- **Central Data Store – DISC Big Data Platform**
  - Application Independent. Common set of analytical tools.
  - Unique Data Repository.
- **Data Integration**
  - Ability to combine data – Enhanced Analytics
  - Single Data Dictionary (**SDD**) + Master Data (**RIAD + CSDB**)

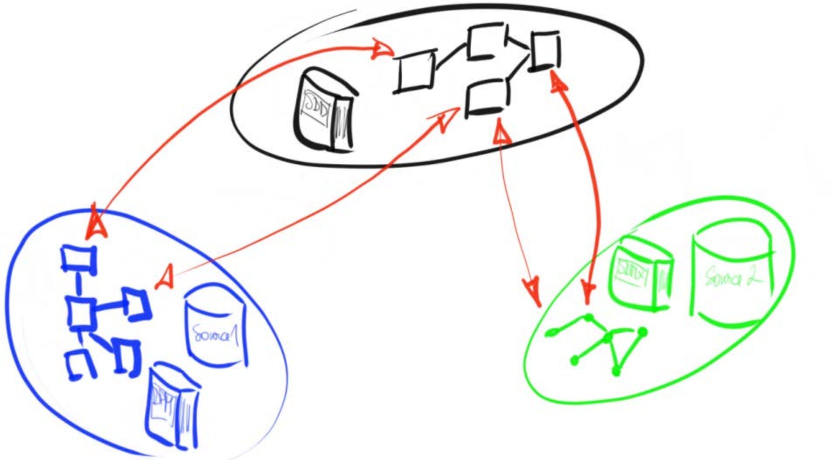


ECB’s approach on Data Integration – SDD

Data integration is the process of ***combining data*** and providing users with a ***unified view*** of the data

The ECB’s approach comprises a ***Single Data Dictionary (SDD)*** that is able to cover the content of other schemas / dictionaries.

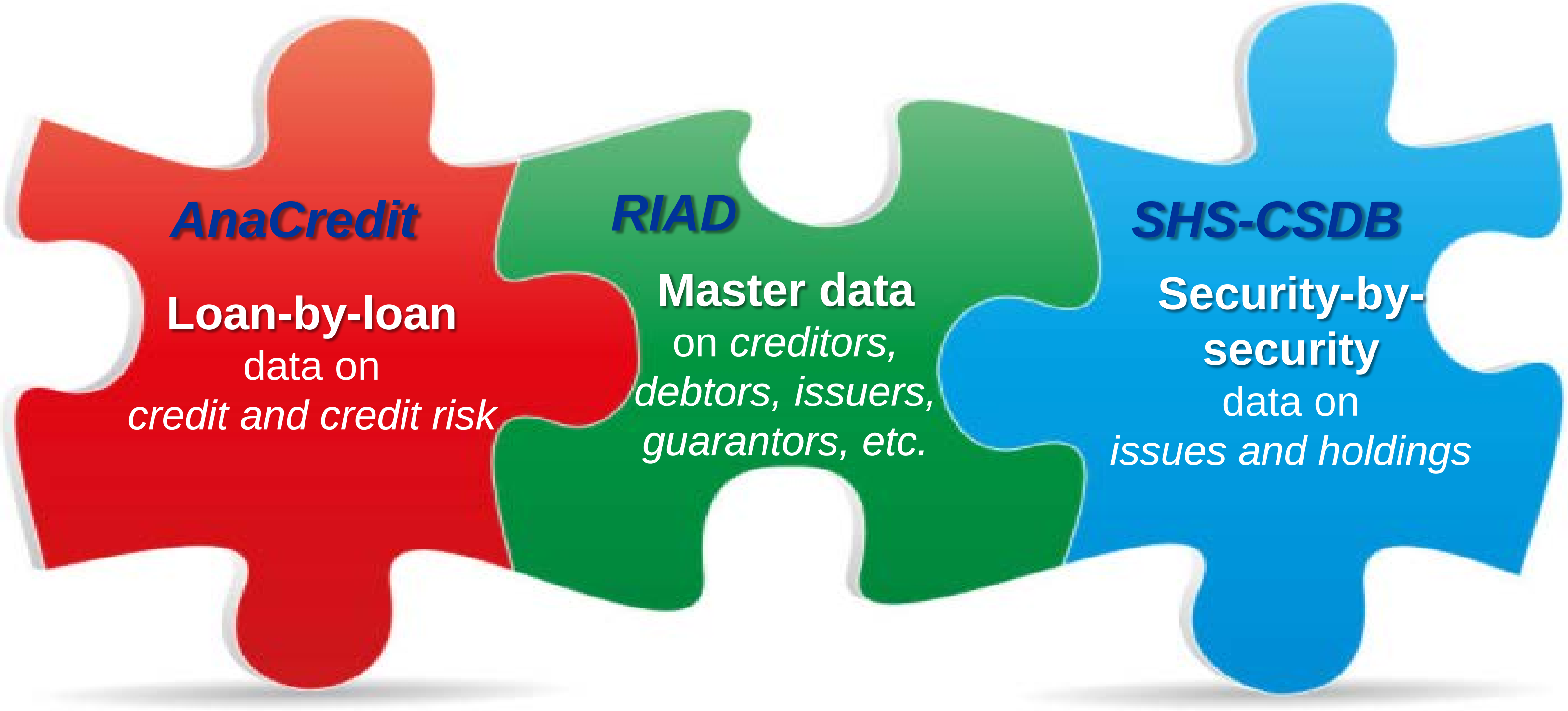




- *Single Data Dictionary (SDD)*
- *Data Point Model (DPM)*
- *Statistical Data and Metadata Exchange (SDMX)*
- *Mappings between dictionaries*

Strategy for data on-boarding and integration - Master Data

RIAD plays a ***pivotal role in integrating*** various granular datasets



RIAD provides ***identification*** of entities and ***relationships between them***, *other datasets* pinpoint actual ***exposures***

RIAD in ***DISC*** is a *prerequisite* to allow integration of granular datasets

Big Data – First experiences - EMIR

EMIR Data on derivatives from six **Trade Repositories**

Daily reporting with **T+1 timeliness**

Both **ETD and OTC** derivatives are reported *since February 2014*

Two-sided reporting obligation (both counterparties to the trade have to submit the report)

EMIR SYSTEM ***since 2017***

	ECB	ESRB
*collected files**	**49,108**	**87,290**
*observations (millions)**	**11,356**	**25,744**
*size of collected files (GB compressed)**	**2,455**	**4,014**

** Data as of 12 December 2018 (the collection started around December 2017)*

Big Data – First experiences - EMIR

ECB-PUBLIC
FINAL

CHALLENGES

SOLUTIONS

IT INFRASTRUCTURE

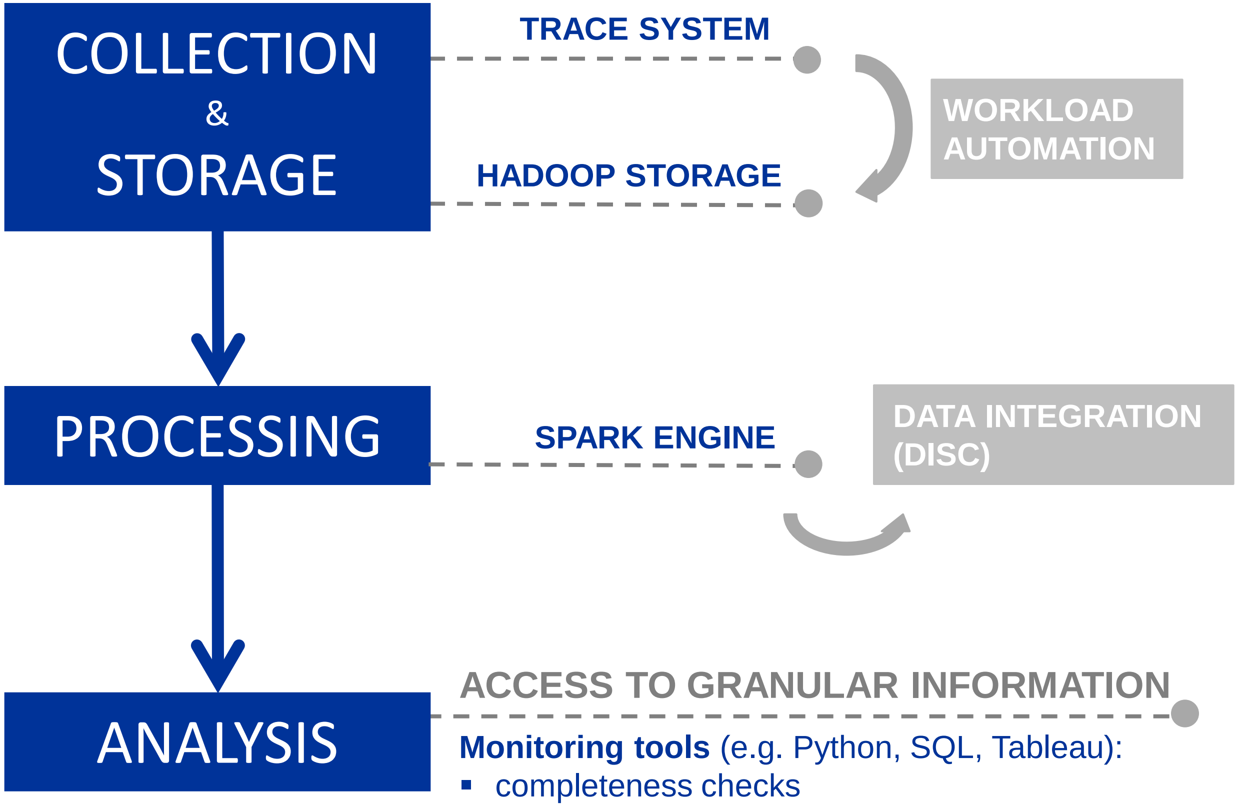
- **Time**: daily frequency
- **Volume**: big data size
- **Volatility**: frequent revisions
- Data Governance & Security

IT INFRASTRUCTURE

- **Time**: processing speed
- **Availability**: EMIR data & other sources (GLEIF, SDW FX, SDW FM, CSDB)
- **Complexity**: double reporting, multiple validation rules

DATA QUALITY

- **Completeness**: coverage
- **Accuracy**: misreporting, enrichment, outliers



COLLECTION & STORAGE

TRACE SYSTEM

HADOOP STORAGE

WORKLOAD AUTOMATION

PROCESSING

SPARK ENGINE

DATA INTEGRATION (DISC)

ANALYSIS

ACCESS TO GRANULAR INFORMATION

Monitoring tools (e.g. Python, SQL, Tableau):

- completeness checks
- quality assessment

Access for final users (Cloudera Hue and CDSW, ODBC)

Big Data – First experiences – SUBA Proof of Concept

ECB-PUBLIC
FINAL

POC with Supervisory Banking (SUBA) data on Hadoop

- Goals:
  - *enable interactive querying on SUBA data*
  - *enable easy data visualization*
  - *assess possibilities and performance of DISC environment*
  - *collect best practices / useful tips*
  - *answer the question: how to best represent SUBA data in Big Data Platform (DISC)?*
- Points of note:
  - *SUBA facts table contains over a billion lines*
  - *SUBA data model is complex, with many tables*
  - *It is similar to the EBA's Data Point Model, with tens of tables, and often requires complicated multi-join queries to get a meaningful and readable result*
- The tools used in this POC:

HIVE
Parquet
python™
cloudera IMPALA
tableau SOFTWARE
APACHE DRILL

POC with SUBA data on Hadoop

- Some conclusions:
  - ***Impala** performs poorly with multi-join queries*
  - *But its speed is impressive when only one huge table is queried*
  - *So… denormalize data with **Hive, Python, Drill**! In order to enhance data locality*
  - *By inserting into the fact table the data related to its foreign keys, we discard the need for joins*
  - *Indeed, when accessing a fact, it is best that relevant features are stored in the same line*
  - *This suits **Parquet** file format nicely: the final table is only 17GB when the initial data was over 150GB when in text format*
  - *It is then possible to connect **Tableau** though ODBC and **Impala** directly on the fact table:*

Machine Learning in DG-Statistics – Use Cases

Typical use cases for Machine Learning (ML)

Large data volumes

Complexity of the data

Ability to identify patterns or relationships that are difficult to detect using statistical modelling

Ability to model expert knowledge in automated way which could improve the timely processing of the data

ML algorithms are computationally intense

Big data platforms – ECB DISC (Hadoop cluster + Cloudera Data Science Workbench)

“Unlimited” storage

High computing power

Parallel processing

Data Science and Machine Learning libraries

Machine Learning in DG-Statistics – Use Cases

Anomaly/Outlier Detection

where standard statistical techniques could not be used

- MMSR
- AnaCredit Outlier Detection and Data Exploration

Data Classification

Assessing, matching or pairing duplicate records

- EMIR
- MMSR

Forecasting, backcasting, interpolating

Estimate missing data using ML algorithms

- Balancing of the Financial Accounts

Record linkage

Link records that represent the same entity in different databases, calibrating missing data by data integration

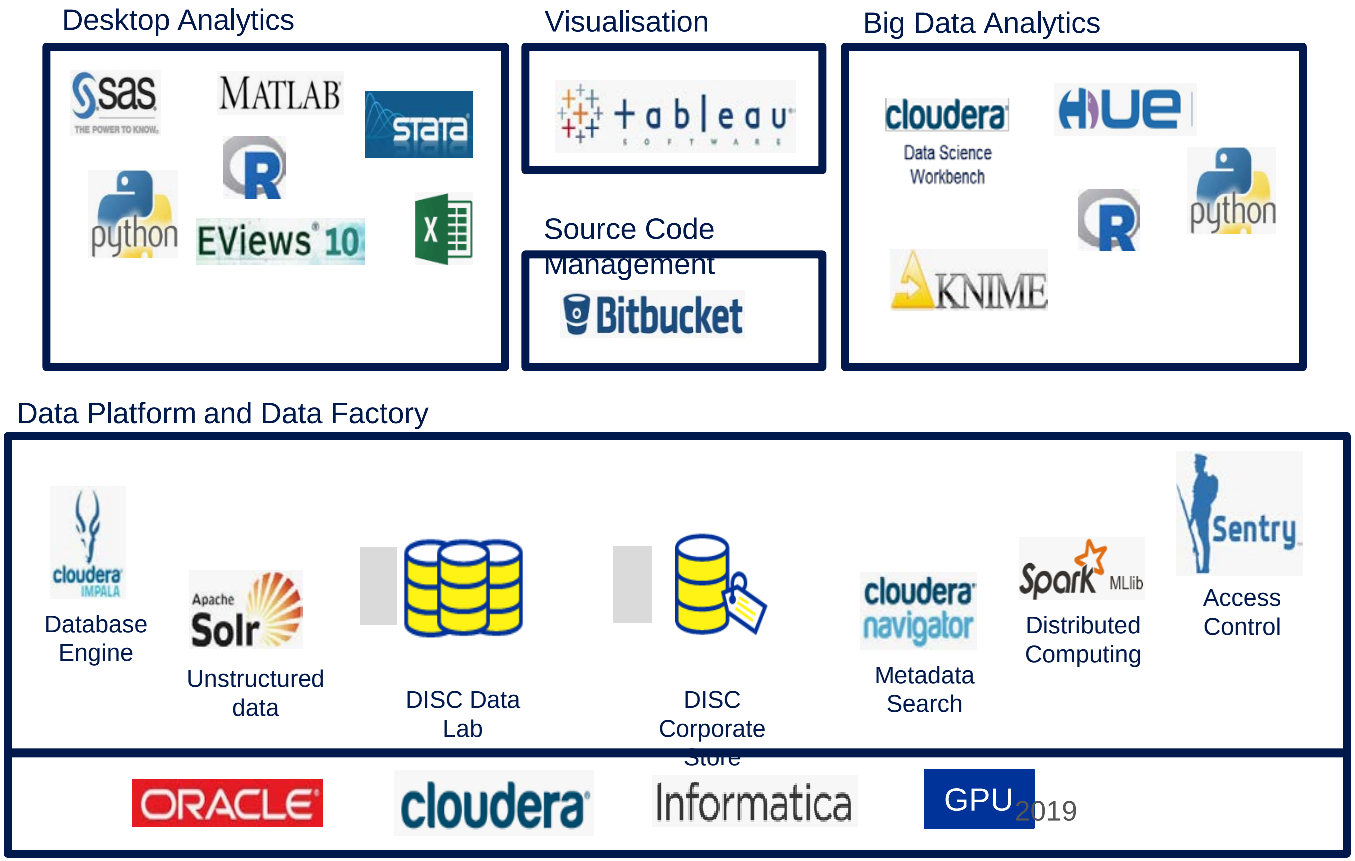
- Institutional sector allocation of MMSR entities based on RIAD

Summary

- Big Data Platform to facilitate analysis
  - Data available in a single platform
  - Integrated datasets → Ability to combine data from different sources
- First outcome with large data sets
  - Positive experiences with EMIR and SUBA
  - New ways of working: models, formats and tools
- Enabling Advanced Analytics (Machine Learning)
  - ECB DISC big data platform - Enabler for ML
  - Data Cleaning
  - Data Classification - Pairing
  - Forecasting
  - Linkage – Missing data

Backup

→ **Variety** of data (structured numerical, structured text, unstructured, web scraping)
→ **Volume** of data, to large to process on single computer (ECB laptop)
→ **Velocity** of changes in data, in particular for unstructured and web scraping use-cases
→ **Know how** to benefit from distributed computing
→ **Find data** and information



Desktop Analytics

Visualisation

Big Data Analytics

Source Code Management

Data Platform and Data Factory

Database Engine

Unstructured data

DISC Data Lab

DISC Corporate Store

Metadata Search

Distributed Computing

Access Control

GPU 2019