Big Data Infrastructure Technologies for Data Analytics and Data Science Projects Operationalisation

Special Session on Big Data

IDAACS2021 22-25 September 2021

Dr. Yuri Demchenko

University of Amsterdam

Outline

- Big Data Reference Architecture and functional components
- Data Science development process
  - Data Science/Data Mining Process models, Dataflow/Data Lifecycle
- Operationalising Data Science Analytics and ML: DataOps, MLOps and platforms
  - Data Science and ML pipeline
- Case study: Data Science and MLOps with Azure
- Further to data driven projects operationalization with Site Reliability Engineering (SRE)
- Summary

Goal of this Research and Talk

- Identify evolution and trends in Big Data technologies and Data Science projects move from desk development to operational stage

- Why it is needed and what is the benefit?

  When trends are identified (among variety of technologies and offers) it is easy to follow and understand further developments in technologies and industrial/business applications

Multiple aspects of Big Data and Relation to Data Science

Big Data is a complex of technologies to enable handling of Big Data (storage, processing, transfer, security)

Data Science is an inter-disciplinary field that uses **scientific methods**, processes, algorithms and systems to extract knowledge and insights from structured and unstructured data.

- Data Science vs Data Analytics vs Statical Analysis

Big Data and Multiple Sources of Data

- Social Media
- Internet
- Communication, voice, video

- Science
- Industrial data
- IoT

- Commonly blending Data Analytics with open and social media data
- IoT is taking a front stage in driving Big Data infrastructure and Data Analytics applications
- Data Spaces demand a full ecosystem with data centric infrastructure and data driven applications

NIST Big Data Working Group (NBD-WG) and ISO/IEC JTC1 Study Group on Big Data (SGBD)

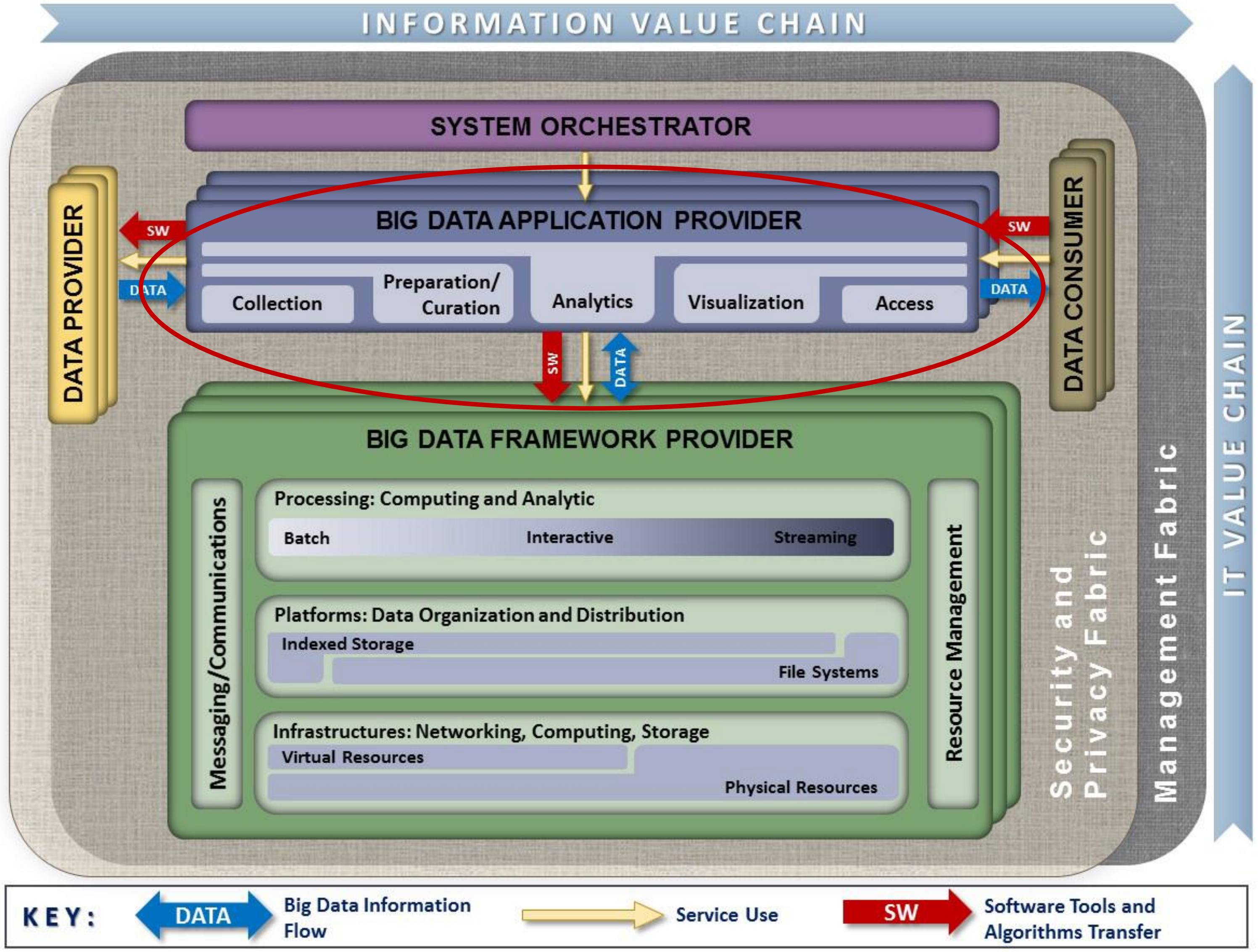
- NIST Big Data Working Group (NBD-WG) is leading the development of the Big Data Technology Roadmap - http://bigdatawg.nist.gov/home.php
  - Built on experience of developing the Cloud Computing standards
- Published as NIST Special Publication 1500 Volumes 1-7 in 2015
- New revision V3 published 2018 - https://bigdatawg.nist.gov/V3_output_docs.php

  Volume 1: **Definitions**
  Volume 2: Taxonomies
  Volume 3: Use Case & Requirements
  Volume 4: **Security & Privacy**
  Volume 5: Architecture White Paper
  Volume 6: **Reference Architecture**
  Volume 7: Standards Roadmap
  Volume 8: **Reference Architecture Interface**
  Volume 9: Modernization and Adoption

The **Big Data Paradigm** consists of the distribution of data systems across horizontally-coupled independent resources to achieve the **scalability** needed for the **efficient processing** of **extensive datasets**.

- NBD-WG defined 3 main components of the new technology:
  - Big Data Paradigm
  - Big Data Science and Data Scientist as a new profession
  - Big Data Architecture

NIST Big Data Reference Architecture (2018)



Main components of the Big Data ecosystem

- Data Provider
- Big Data Applications Provider
- Big Data Framework Provider
- Data Consumer
- Service Orchestrator

Big Data Lifecycle and Applications Provider activities

- Collection
- Preparation
- Analysis and Analytics
- Visualization
- Access

Big Data Ecosystem includes all components that are involved in the big data production, processing, delivery, and consuming

[ref] Volume 6: NIST Big Data Reference Architecture. http://bigdatawg.nist.gov/V1_output_docs.php

CCI and UvA Research on Big Data Technologies and Scientific Data Infrastructure – Journey driven by Technology Development

- 2010-2015 Starting Cloud Computing research and contribution to NIST Big Data WG on SP 1500 first edition 2015
- Numerous projects related to e-Infrastructure and European Research Infrastructures
  - EGEE, GEANT, EOSC FAIR, ENVRI, SLICES-RI, AMDeX
- Research on Big Data Infrastructure technologies and data centric applications
  - Platform Engineering and adopting Digital Platform Architecture model
- Data Spaces and IDSA Reference Architecture, Interoperability Framework
  - Data Spaces optimization for Trustworthiness, Sustainability, Societally Responsible Development
  - Data sharing and Data Exchanges
- DataOps: DevOps for Data Science projects – Operationalisation of Data Science projects

Data Science Project Management

- Data Science development process
- Data Science/Data Mining Process models, Data Flow/Data Lifecycle
- Operationalising Data Science Analytics and ML: DataOps, MLOps and platforms
- Case study: Azure Analytics stack and MLOps platform

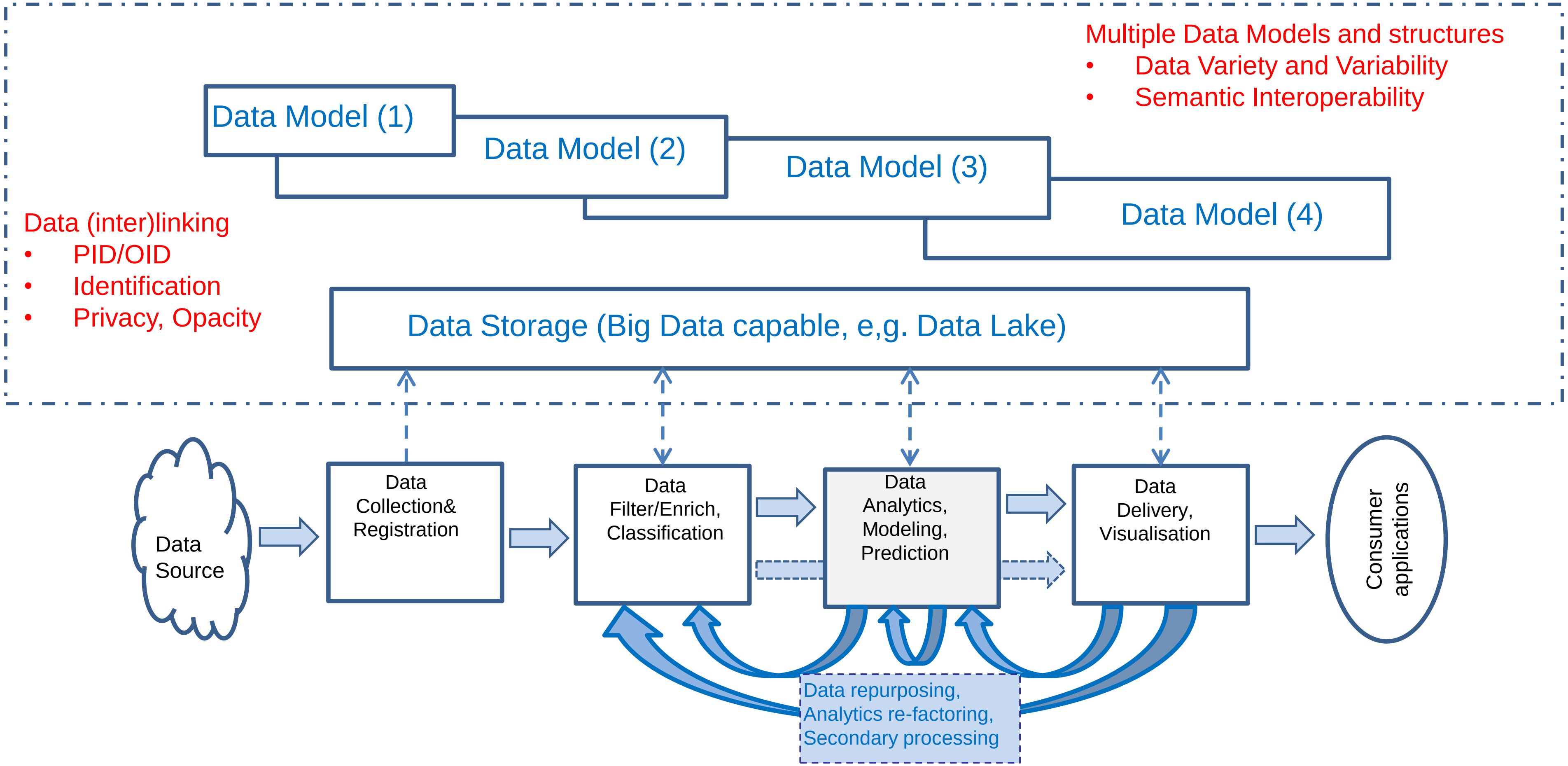
Data Science Development Process

Data Science process is dealing with the data pipelines that include stages:

- **Collecting data** from multiple sources, also blending process or business data with external data such as environmental data or social media data that can be obtained via WebAPI or web scraping
- **Working with data** including data preparation, cleaning, filtering, and reformatting for modeling needs
- **Combing datasets** by joining on common attributes, consolidating attributes, build tabular data structure (such as used in popular analytics programming languages R, python, scala)
- **Feature engineering, algorithm selection**
- **Testing** before production and continuously **validating** the model during production, in particular, detecting drift in predictive models
- Implement changes and **deploy an updated model**.

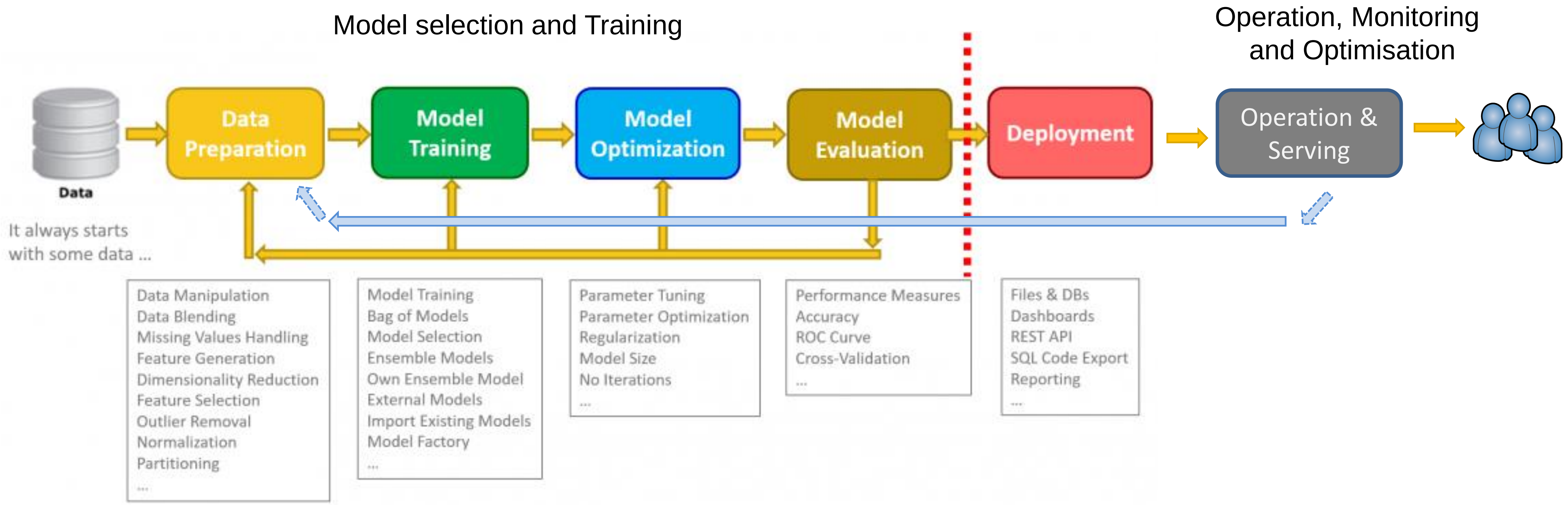
- Data Science process is linked to Data Lifecycle and Dataflow

Data Lifecycle/Transformation Model


Multiple Data Models and structures
• Data Variety and Variability
• Semantic Interoperability
Data Model (1)
Data Model (2)
Data Model (3)
Data Model (4)
Data (inter)linking
• PID/OID
• Identification
• Privacy, Opacity
Data Storage (Big Data capable, e,g. Data Lake)
Data Source
Data Collection& Registration
Data Filter/Enrich, Classification
Data Analytics, Modeling, Prediction
Data Delivery, Visualisation
Consumer applications
Data repurposing, Analytics re-factoring, Secondary processing

- Data Model changes along data lifecycle or evolution (Variability)
- Data provenance (lineage) is a discipline to track all data transformations along their lifecycle
- Identifying and linking data
  - Persistent data/object identifiers (PID/OID)
  - Traceability vs Opacity
  - Referral integrity

Data Science Project/Research is based on Dataflow/Lifecycle



Model selection and Training

Operation, Monitoring and Optimisation

Data

It always starts with some data ...

Data Preparation

Model Training

Model Optimization

Model Evaluation

Deployment

Operation & Serving

Data Manipulation
Data Blending
Missing Values Handling
Feature Generation
Dimensionality Reduction
Feature Selection
Outlier Removal
Normalization
Partitioning
...

Model Training
Bag of Models
Model Selection
Ensemble Models
Own Ensemble Model
External Models
Import Existing Models
Model Factory
...

Parameter Tuning
Parameter Optimization
Regularization
Model Size
No Iterations
...

Performance Measures
Accuracy
ROC Curve
Cross-Validation
...

Files & DBs
Dashboards
REST API
SQL Code Export
Reporting
...

- **Each phase** – data preparation, model training and evaluation, and model deployment – operates on **its own data set**. All these data sets need to be isolated but linked. The pollution of data sets across the data science assembly line is one of the most frequent mistakes in model production.
- The data science project is typically starts with some ***historical data*** *or* ***sample dataset*** can be somewhere in existing repositories.
- Data preparation may also include connecting external data sources – ***data blending***

[Ref] https://www.knime.com/blog/analytics-and-beyond

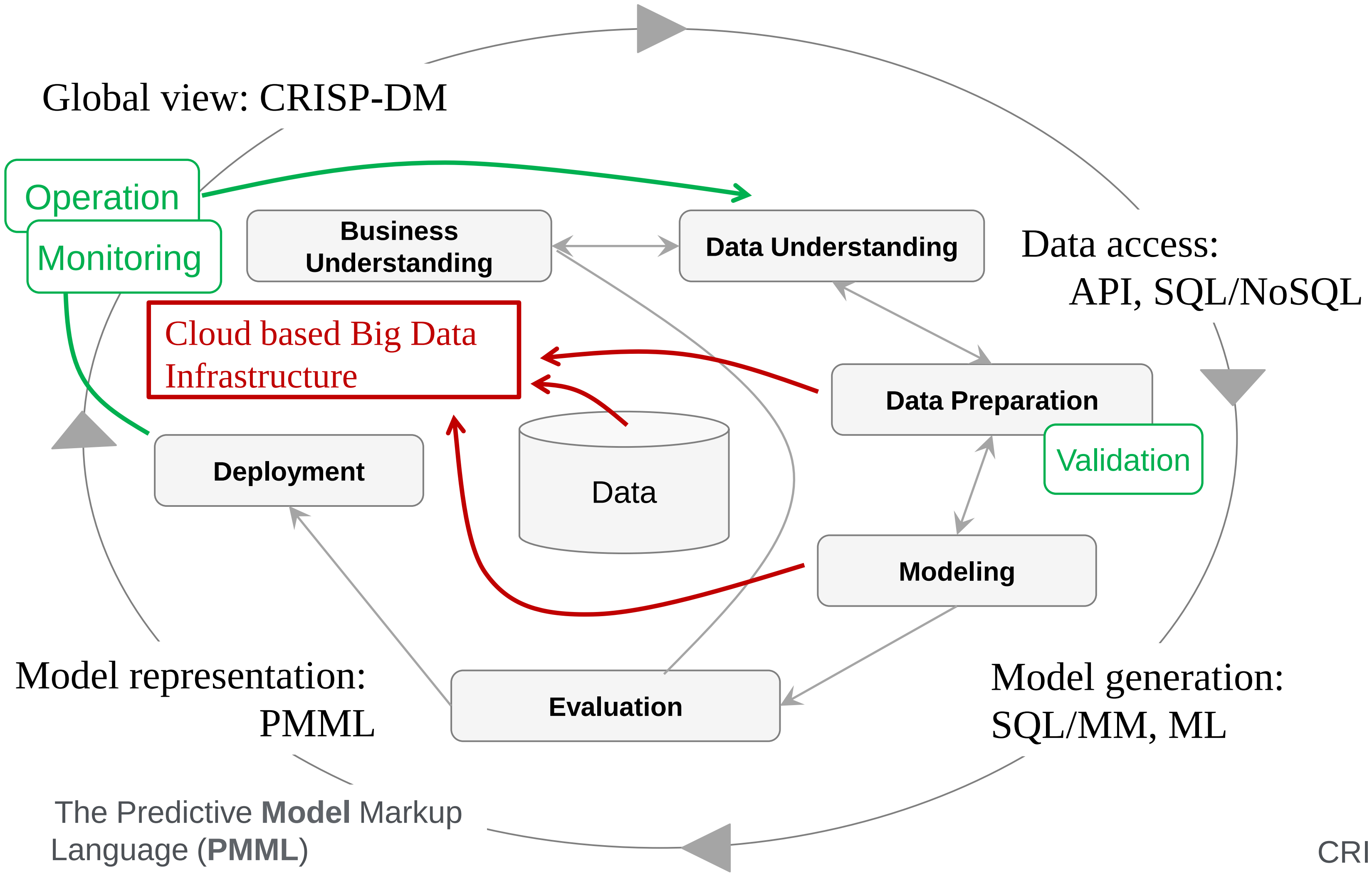
Data Science Process Models and Model Formats

- Data Science process models
  - CRISP-DM, CRoss-Industry Standard Process for Data Mining
  - ASUM, Analytics Solutions Unified Method (IBM)
  - TDSP, Team Data Science Process (Microsoft)
  - KNIME Model Factory (KMF)

- Data Analytics Model Formats
  - Predictive Models Markup Language (PMML)
  - Portable Format for Analytics (PFA)
  - ONNX (Open Neural Network Exchange)
  - TensorFlow Model

From model creation to deployment on Big Data/cloud platform to Production

CRISP DM Process: Processes and Data Lifecycle
(Cross-Industry Standard Process for Data Mining)



The Predictive **Model** Markup Language (**PMML**)

Cross Industry Standard Process for Data Mining (CRISP-DM) model and stages

- Business understanding
- Data Understanding
- Data preparation
  - Data Validation
- Modelling
- Evaluation
- Deployment => Operation
  - Process monitoring

All stages are iterative with the goal to achieve effectiveness for business decision making

CRISP-DM historically published 1st version in 1999

MLOps and DataOps: DevOps for ML and Data Analytics

MLOps and DataOps are extension of **DevOps** to manage Data Analytics and Machine Learning data flow and process.

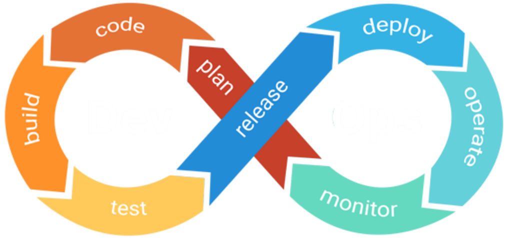
- DevOps is the practice of operations and development engineers participating together in the entire service lifecycle, from design through the development process to production support.
- Develop – Build – Deploy – Operate
- Cloud is an enabler for DevOps processes

DataOps, MLOps is about operationalizing ML and Data Analytics

- *Different nature of processes at the stage of ML model development*
- Benefit from the DevOps processes and culture
- Learn from DevOps experience

DevOps Essentials (from Software Engineering)

- Better Software, Faster time to market
- Synergy of Development and Operation
- Covers the *entire* Application Lifecycle
- Continuous Improvement in CI/CD cycle

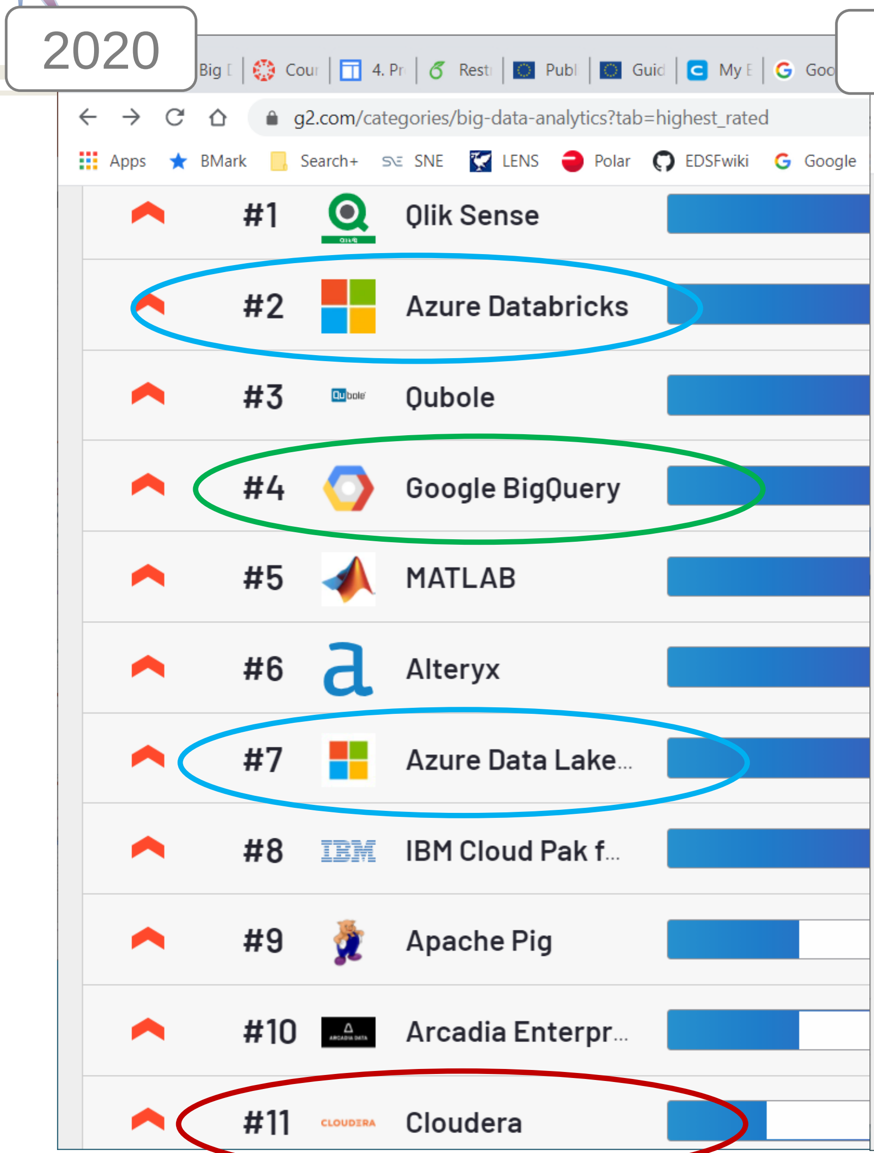


Big Data Platform: Research, Development and Production

- Big Data require cloud scale due to Big Data data gravity: Compute, Storage and Network optimization + Energy consumption
- All major cloud providers offer Big Data platforms and services
  - Amazon Web Services (AWS)
  - Microsoft Azure
  - Google Cloud Platform (GCP)
- Specialised Data Analytics providers and vendors

The Top 11 (2020) and Top 14 (2021) Big Data Analytics Software [ref]

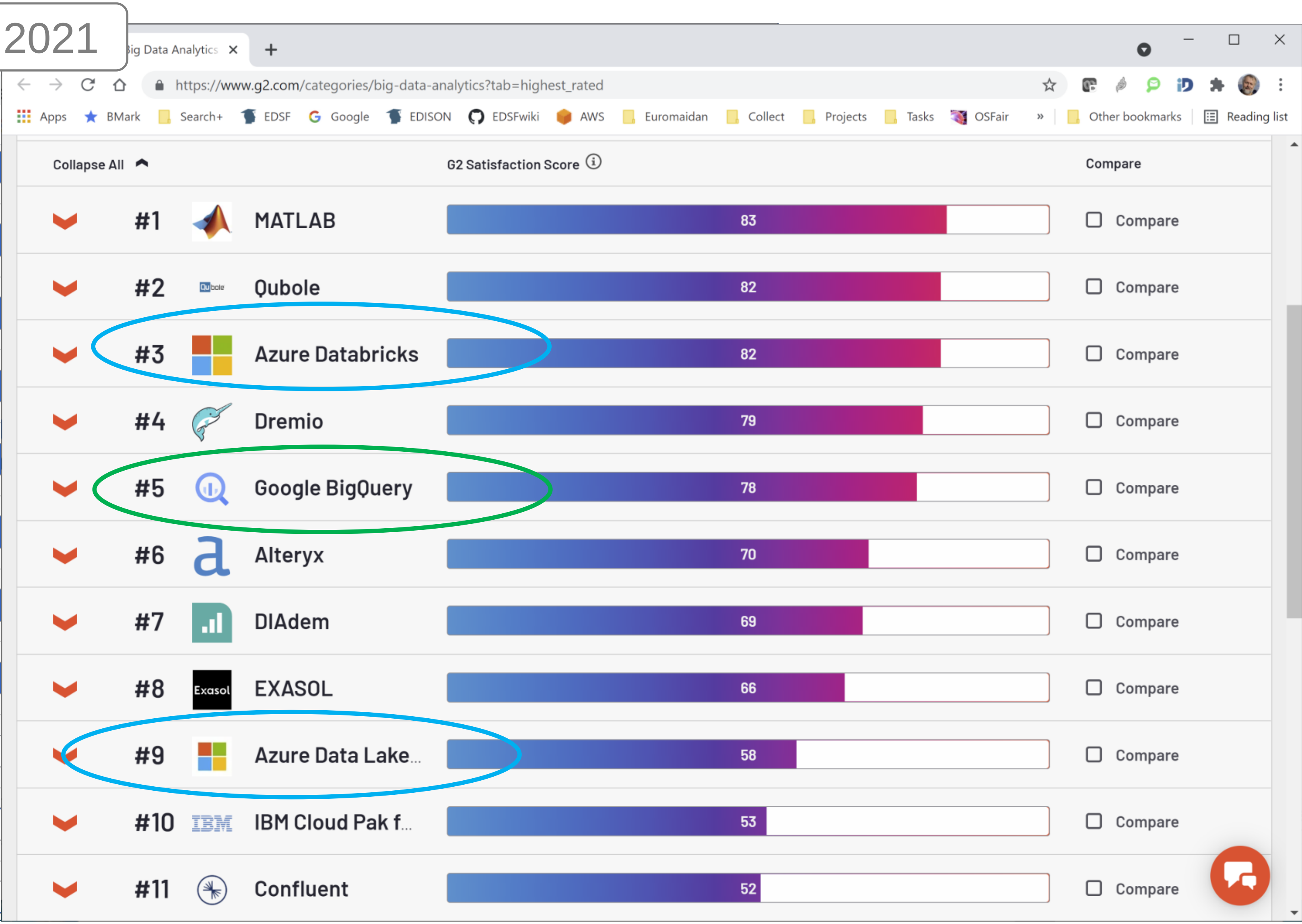
2020



g2.com/categories/big-data-analytics?tab=highest_rated

#1	Qlik Sense
#2	Azure Databricks
#3	Qubole
#4	Google BigQuery
#5	MATLAB
#6	Alteryx
#7	Azure Data Lake...
#8	IBM Cloud Pak f...
#9	Apache Pig
#10	Arcadia Enterpr...
#11	Cloudera

2021

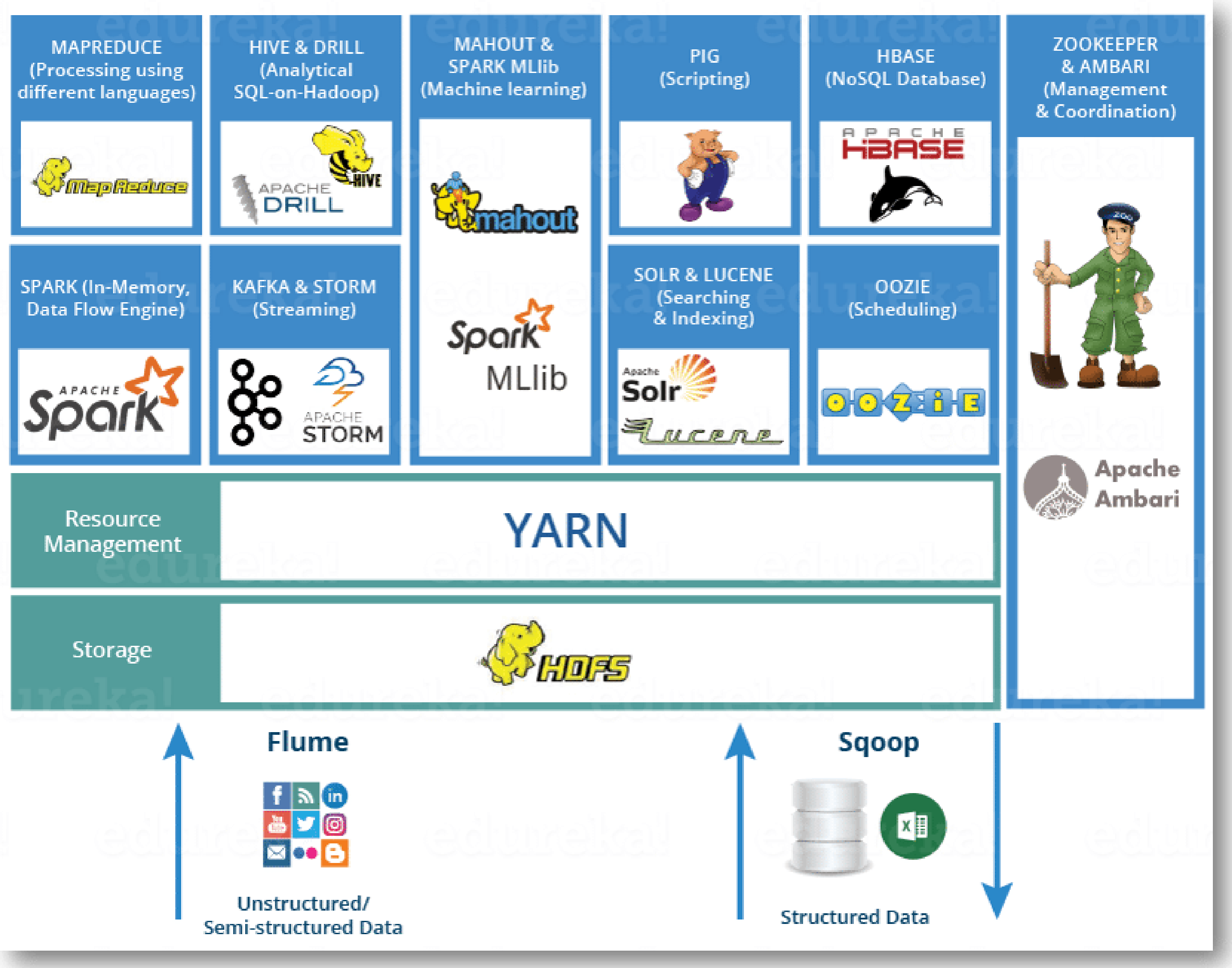


https://www.g2.com/categories/big-data-analytics?tab=highest_rated

Collapse All		G2 Satisfaction Score	Compare
#1	MATLAB	83	□ Compare
#2	Qubole	82	□ Compare
#3	Azure Databricks	82	□ Compare
#4	Dremio	79	□ Compare
#5	Google BigQuery	78	□ Compare
#6	Alteryx	70	□ Compare
#7	DIAdem	69	□ Compare
#8	EXASOL	66	□ Compare
#9	Azure Data Lake...	58	□ Compare
#10	IBM Cloud Pak f...	53	□ Compare
#11	Confluent	52	□ Compare

[ref] https://www.g2.com/categories/big-data-analytics?tab=highest_rated

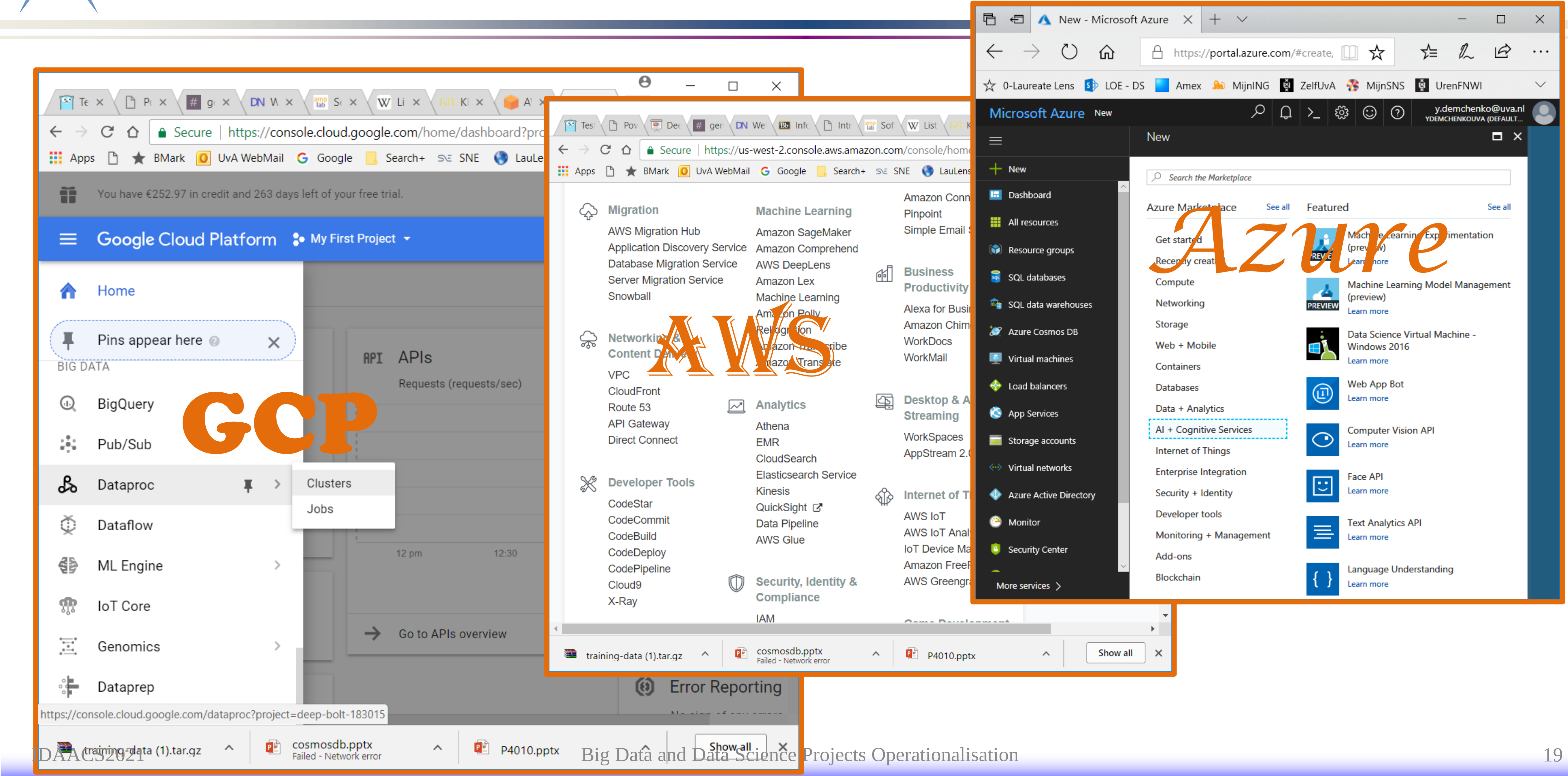
Hadoop Ecosystem: Platform for Big Data Analytics


MAPREDUCE (Processing using different languages)
MapReduce
HIVE & DRILL (Analytical SQL-on-Hadoop)
APACHE DRILL
HIVE
MAHOUT & SPARK MLlib (Machine learning)
mahout
Spark MLlib
PIG (Scripting)
HBASE (NoSQL Database)
APACHE HBASE
ZOOKEEPER & AMBARI (Management & Coordination)
Apache Ambari
SPARK (In-Memory, Data Flow Engine)
APACHE Spark
KAFKA & STORM (Streaming)
APACHE STORM
SOLR & LUCENE (Searching & Indexing)
Apache Solr
Lucene
OOZIE (Scheduling)
OOZIE
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/ Semi-structured Data
Sqoop
Structured Data
edureka!

- Basically Open Source with multiple commercial service platforms
- Amazon EMR – Elastic Map Reduce
  - Includes Spark
- Azure HDInsight
- Azure Databricks and Delta Lake (from Databricks)
- Cloudera/Hortonworks

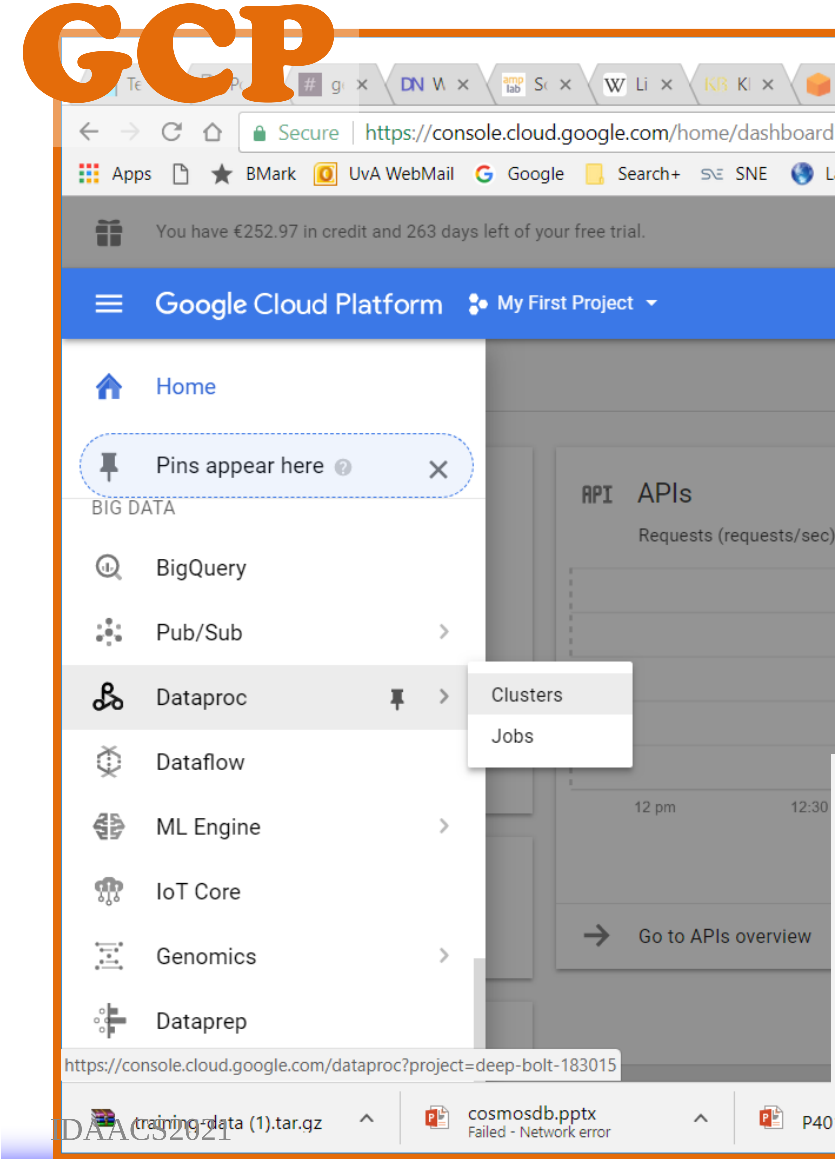
- Data Lakes technology is based on HDFS (Hadoop Distributed File System)

Google, AWS, Azure Big Data Stacks

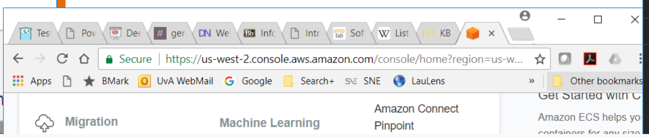


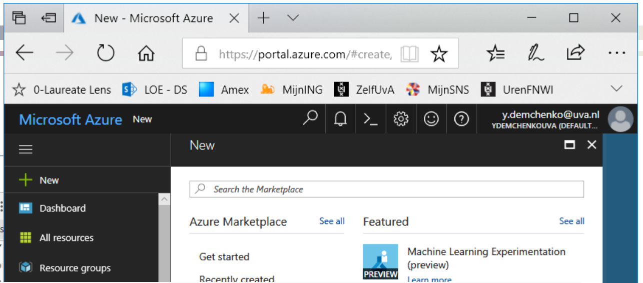
Google Cloud Platform

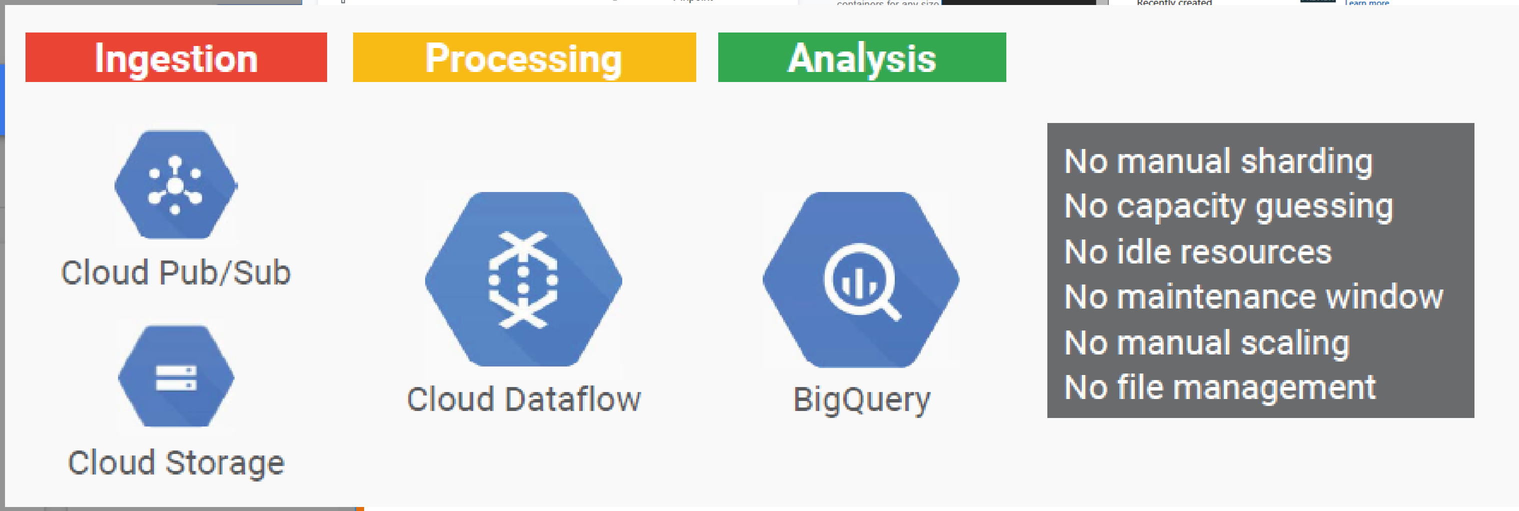
GCP



You have €252.97 in credit and 263 days left of your free trial.



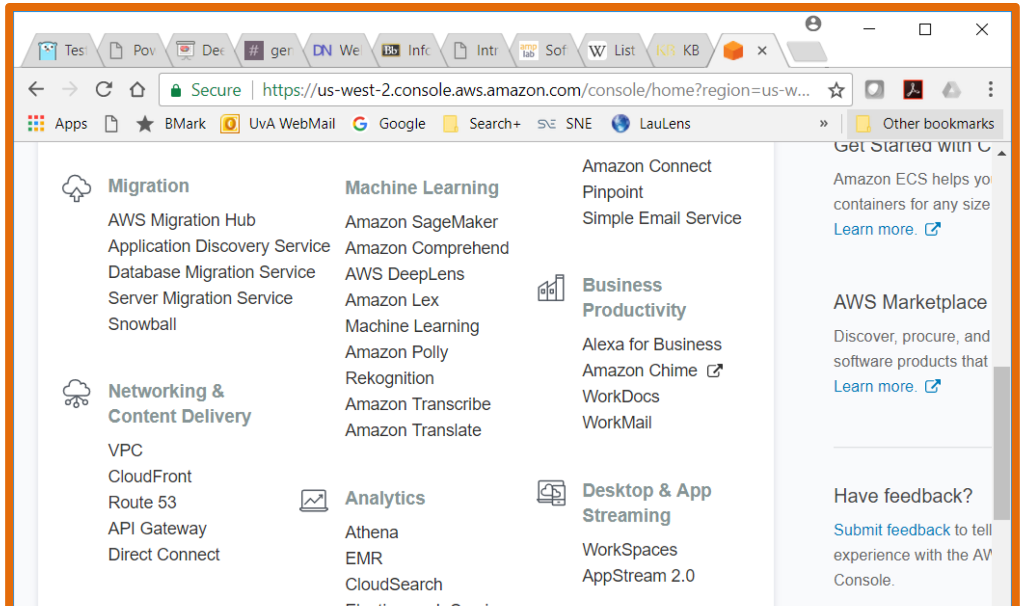




No manual sharding
No capacity guessing
No idle resources
No maintenance window
No manual scaling
No file management

- Machine Learning embedded across most products
- Multiple Tensorflow ML models in use
  - Portable TensorFlow models
- Key models exposed via APIs (Democratizing Machine Learning)
  - Cloud Video Intelligence API
  - Cloud Vision API
  - Cloud Natural Language API
  - Cloud Translation API
  - Cloud Speech API
- Acquired Kaggle in 2017 - Data Science Enthusiasts

Amazon Web Services (AWS)



Tes | Pov | Dec | ger | DN We | Bb Info | Intr | Sof | W List | KB

Secure | https://us-west-2.console.aws.amazon.com/console/home?region=us-w...

Apps | BMark | UvA WebMail | Google | Search+ | SNE | LauLens | » | Other bookmarks

Migration
AWS Migration Hub
Application Discovery Service
Database Migration Service
Server Migration Service
Snowball

Networking & Content Delivery
VPC
CloudFront
Route 53
API Gateway
Direct Connect

Developer Tools
CodeStar
CodeCommit
CodeBuild
CodeDeploy
CodePipeline
Cloud9
X-Ray

Machine Learning
Amazon SageMaker
Amazon Comprehend
AWS DeepLens
Amazon Lex
Machine Learning
Amazon Polly
Rekognition
Amazon Transcribe
Amazon Translate

Analytics
Athena
EMR
CloudSearch
Elasticsearch Service
Kinesis
QuickSight
Data Pipeline
AWS Glue

Security, Identity & Compliance
IAM

Amazon Connect
Pinpoint
Simple Email Service

Business Productivity
Alexa for Business
Amazon Chime
WorkDocs
WorkMail

Desktop & App Streaming
WorkSpaces
AppStream 2.0

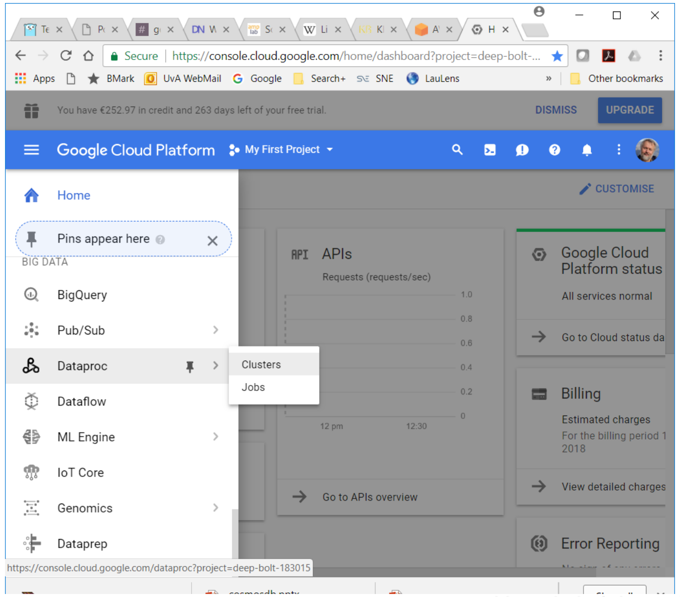
Inter
AWS
AWS
IoT D
Ama
AWS

Get Started with C
Amazon ECS helps yo containers for any size
Learn more.

AWS Marketplace
Discover, procure, and software products that
Learn more.

Have feedback?
Submit feedback to tell experience with the AW Console.

training-data (1).tar.gz | cosmosdb.pptx Failed - Network error



Te | P | g | DN W | S | W Li | K | A | H

Secure | https://console.cloud.google.com/home/dashboard?project=deep-bolt-...

Apps | BMark | UvA WebMail | Google | Search+ | SNE | LauLens | » | Other bookmarks

You have €252.97 in credit and 263 days left of your free trial. DISMISS UPGRADE

Google Cloud Platform | My First Project

Home

Pins appear here

BIG DATA
- BigQuery
- Pub/Sub
- Dataproc — Clusters, Jobs
- Dataflow
- ML Engine
- IoT Core
- Genomics
- Dataprep

CUSTOMISE

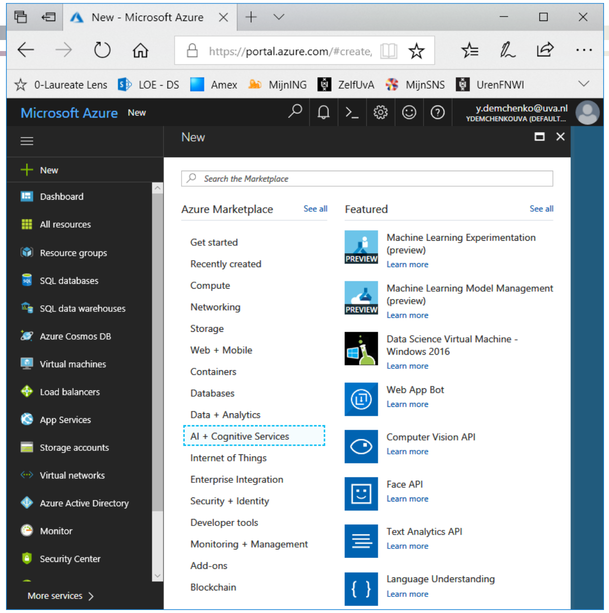
APIs
Requests (requests/sec)
1.0 0.8 0.6 0.4 0.2 0
12 pm 12:30
Go to APIs overview

Google Cloud Platform status
All services normal
Go to Cloud status da

Billing
Estimated charges
For the billing period 1 2018
View detailed charges

Error Reporting

https://console.cloud.google.com/dataproc?project=deep-bolt-183015



New - Microsoft Azure

https://portal.azure.com/#create,

0-Laureate Lens | LOE - DS | Amex | MijnING | ZelfUvA | MijnSNS | UrenFNWI

Microsoft Azure New | y.demchenko@uva.nl YDEMCHENKOUVA (DEFAULT...

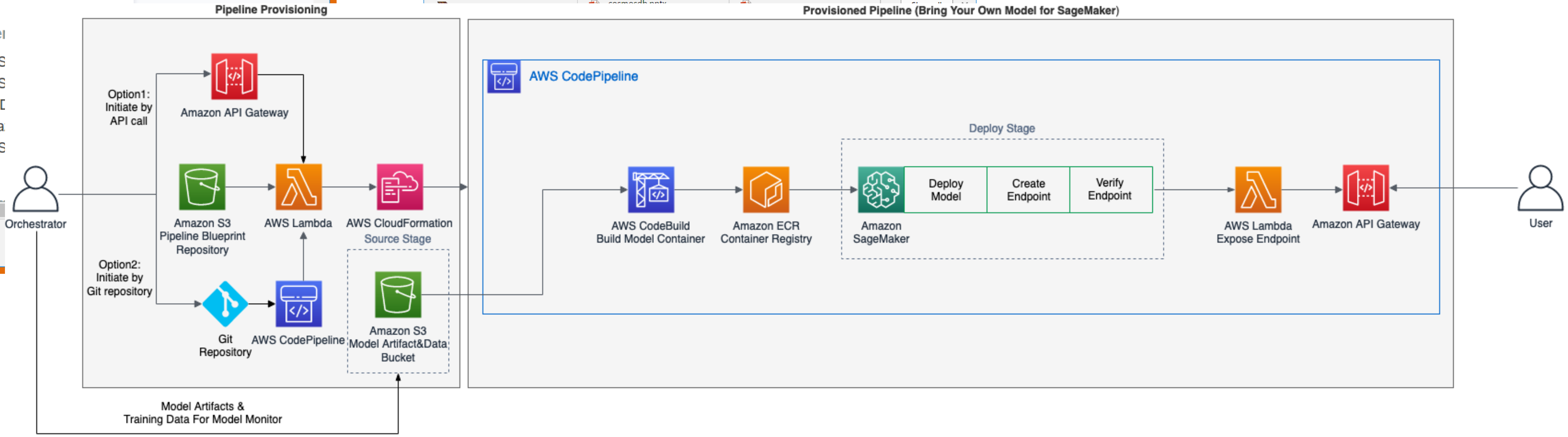
- New
- Dashboard
- All resources
- Resource groups
- SQL databases
- SQL data warehouses
- Azure Cosmos DB
- Virtual machines
- Load balancers
- App Services
- Storage accounts
- Virtual networks
- Azure Active Directory
- Monitor
- Security Center
- More services >

New

Search the Marketplace

Azure Marketplace — See all
- Get started
- Recently created
- Compute
- Networking
- Storage
- Web + Mobile
- Containers
- Databases
- Data + Analytics
- AI + Cognitive Services
- Internet of Things
- Enterprise Integration
- Security + Identity
- Developer tools
- Monitoring + Management
- Add-ons
- Blockchain

Featured — See all
- Machine Learning Experimentation (preview) — Learn more
- Machine Learning Model Management (preview) — Learn more
- Data Science Virtual Machine - Windows 2016 — Learn more
- Web App Bot — Learn more
- Computer Vision API — Learn more
- Face API — Learn more
- Text Analytics API — Learn more
- Language Understanding — Learn more



Pipeline Provisioning
Option1: Initiate by API call
Amazon API Gateway
Orchestrator
Amazon S3 Pipeline Blueprint Repository
AWS Lambda
AWS CloudFormation
Source Stage
Option2: Initiate by Git repository
Git Repository
AWS CodePipeline
Amazon S3 Model Artifact&Data Bucket
Model Artifacts & Training Data For Model Monitor
Provisioned Pipeline (Bring Your Own Model for SageMaker)
AWS CodePipeline
Deploy Stage
AWS CodeBuild Build Model Container
Amazon ECR Container Registry
Amazon SageMaker
Deploy Model
Create Endpoint
Verify Endpoint
AWS Lambda Expose Endpoint
Amazon API Gateway
User

AWS

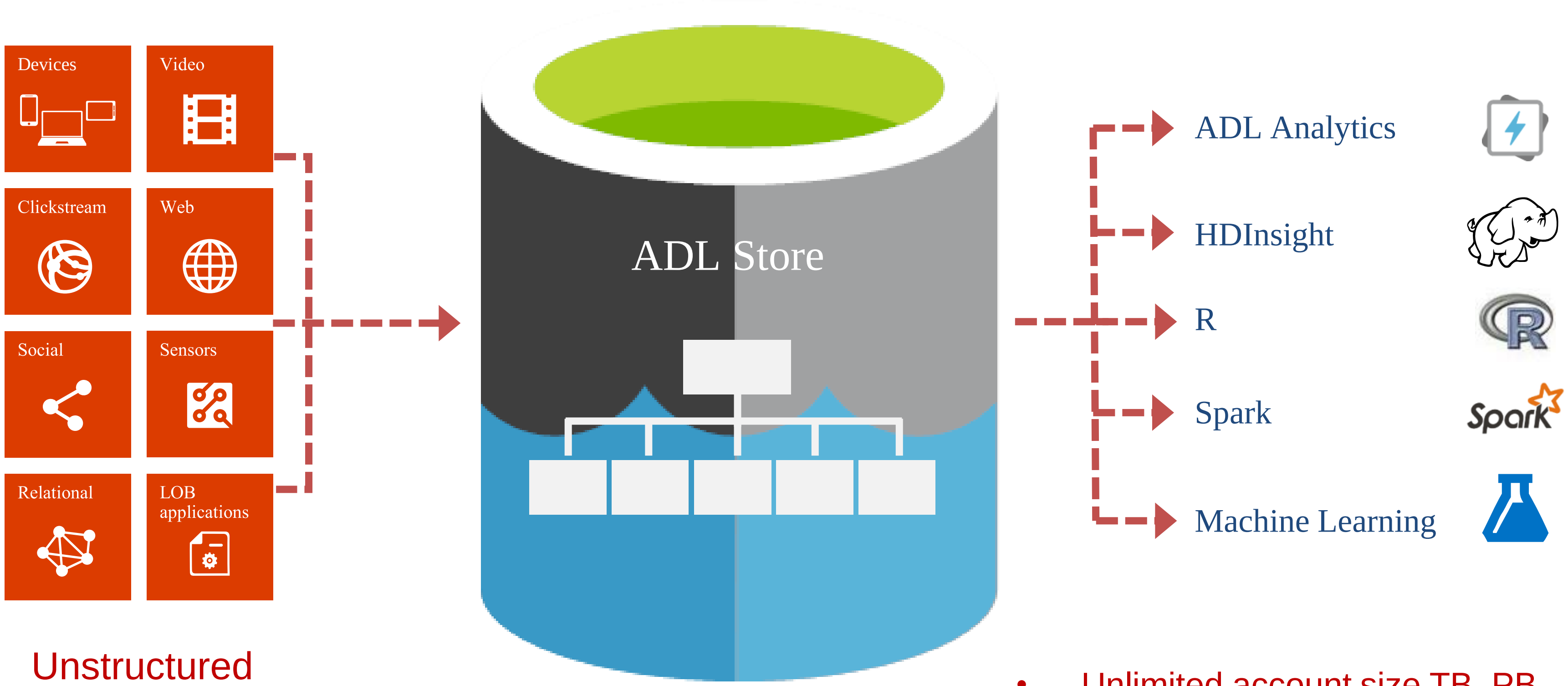
Case Study: Azure Data Science and Analytics Stack

- Empowering Data Science Process with DevOps and MLOps
- Business oriented with good integration with Biz Analytics platforms

- Azure Data Lake and Delta Lake
- Azure HDInsight and Azure Databricks Spark
- Azure DevOps and MLOps

What is Azure Data Lake (ADL) Store?

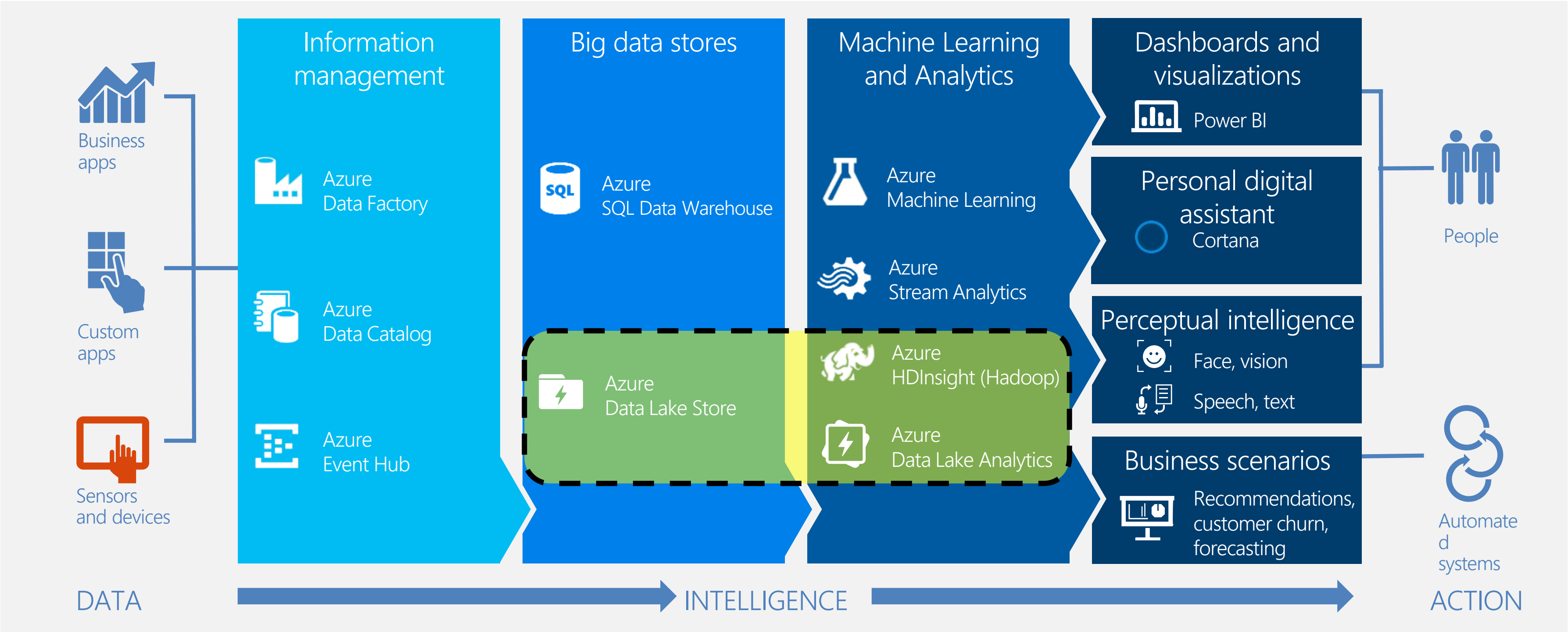
A highly scalable, distributed, parallel file system in the cloud specifically designed to work with multiple analytic frameworks



- Unstructured
- Semi-structured
- Structured

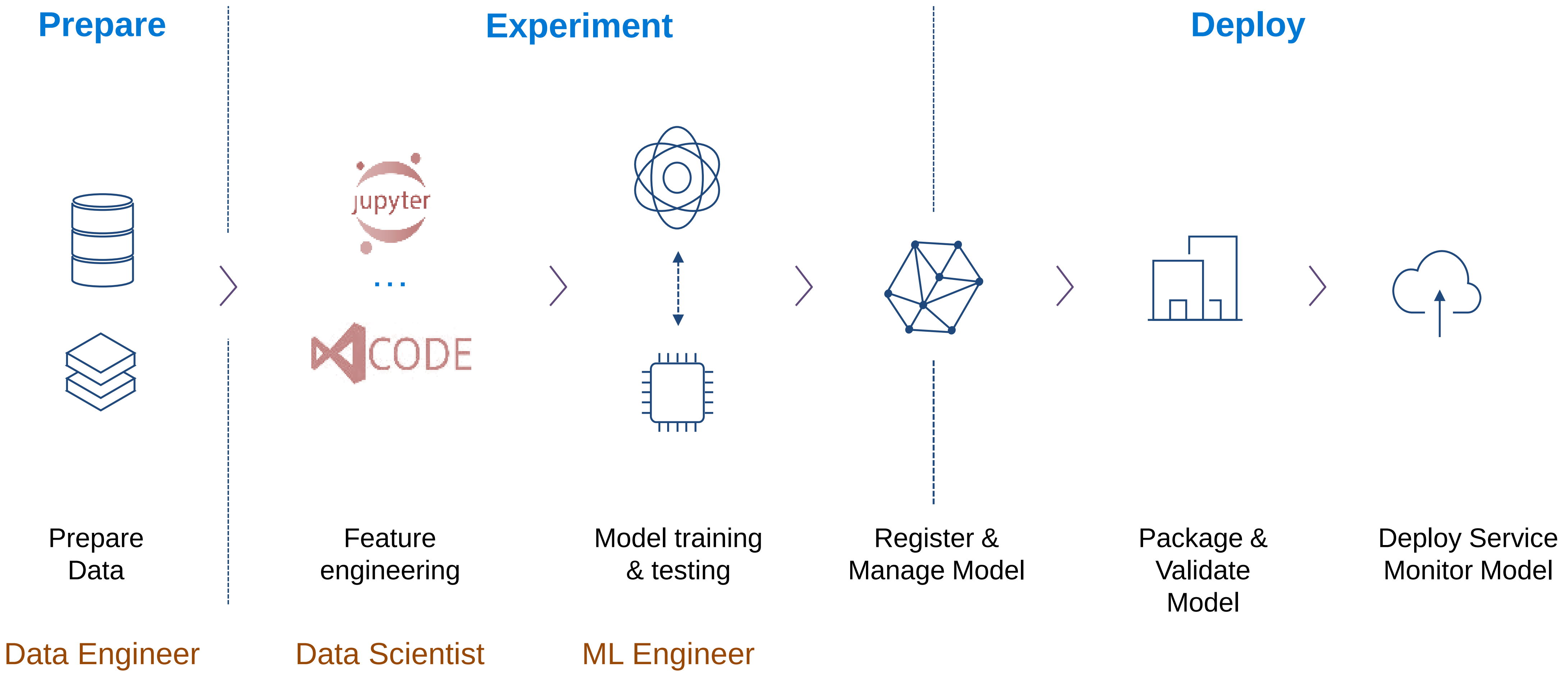
- Unlimited account size TB, PB
- Individual files size from gigabytes to petabytes
- No limits to scale

Azure Data Lake: Part of Cortana Analytics Suite



- Azure Stack for Azure services deployment at premises
- Fully hybrid Data Lake Analytics
- Automatic load balancing and outsourcing to Azure cloud

Azure MLOps: Data Science Workflow



Prepare | **Experiment** | **Deploy**

Prepare Data

Feature engineering

Model training & testing

Register & Manage Model

Package & Validate Model

Deploy Service Monitor Model

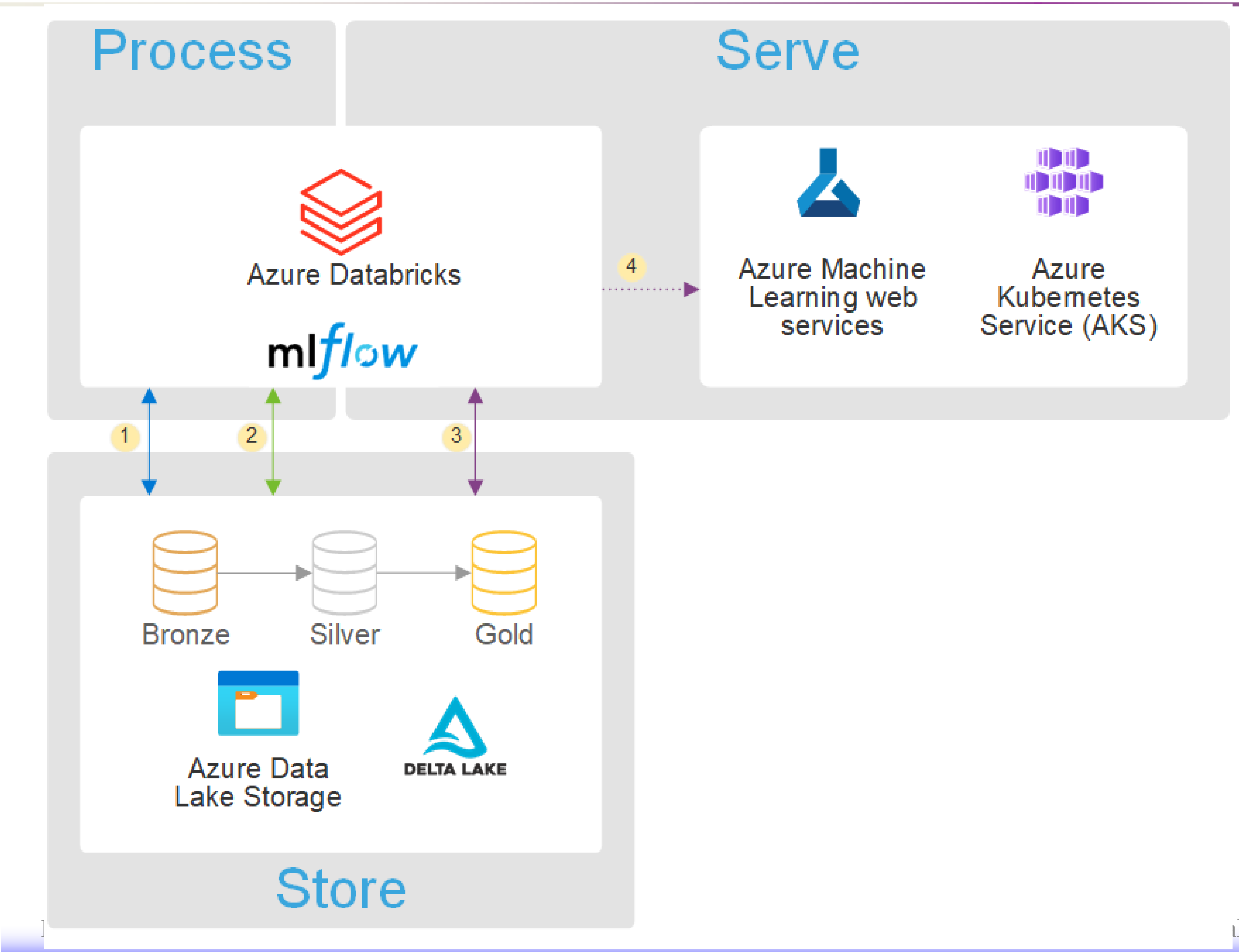
Data Engineer

Data Scientist

ML Engineer

Azure Data Science and Machine Learning Environment

https://docs.microsoft.com/en-us/azure/architecture/solution-ideas/articles/azure-databricks-data-science-machine-learning

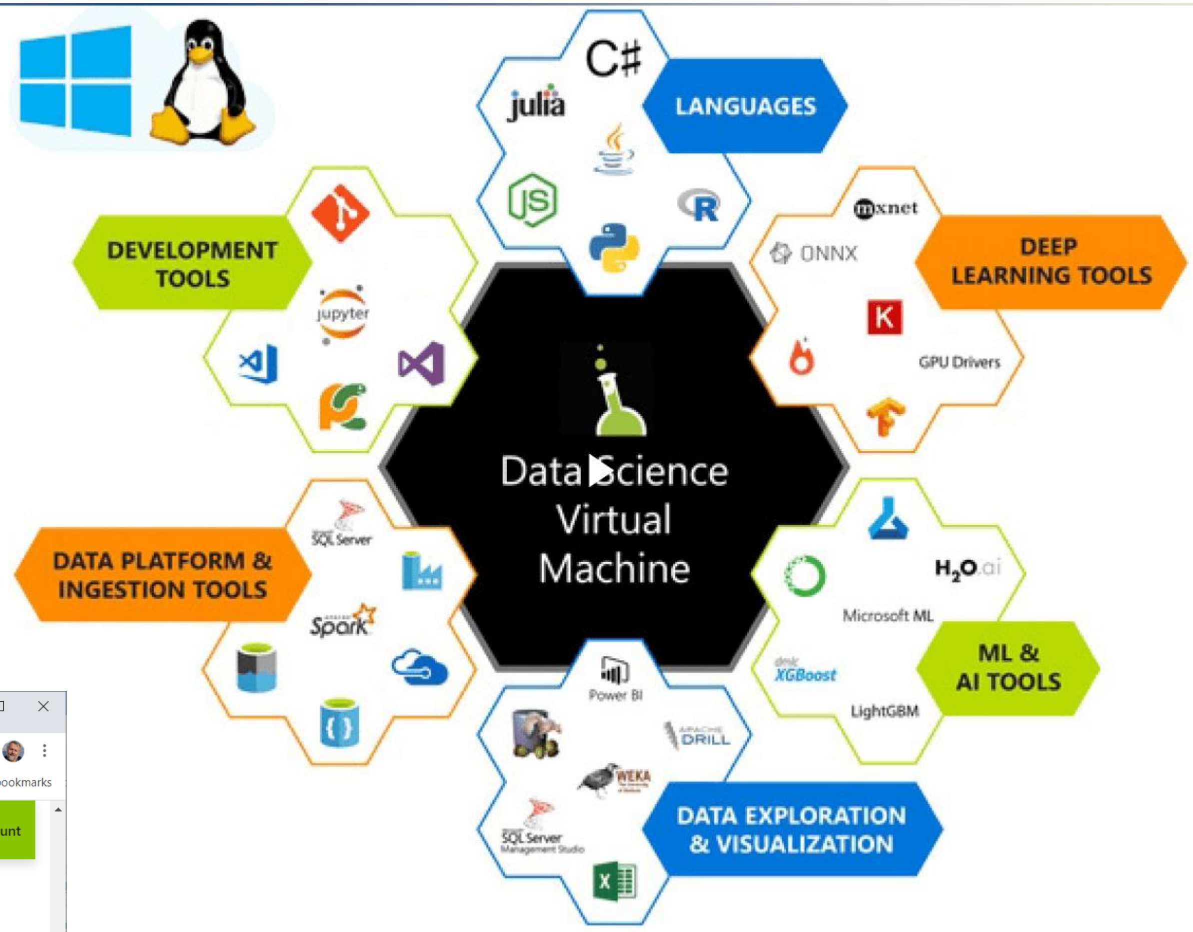


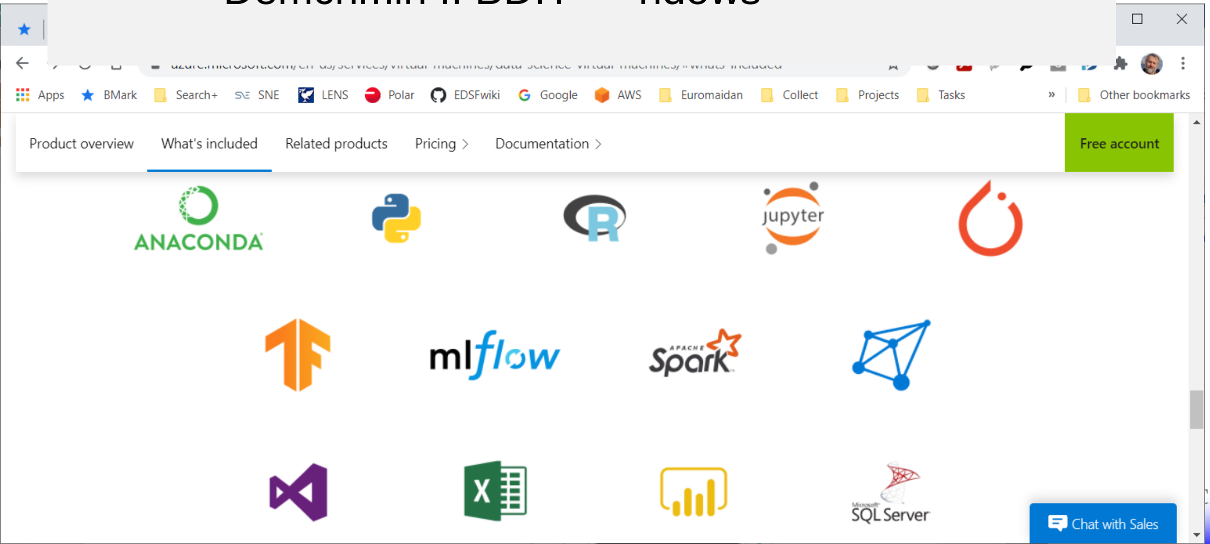
- Optimized Spark engine
- Machine learning run time
  - PyTorch, TensorFlow, and scikit-learn
- MLflow
- Choice of language
  - Python, Scala, R, Spark SQL and .Net
- Collaborative notebooks
- Delta lake
- Native integrations with Azure services
- Interactive workspaces
- Enterprise-grade security
- Production-ready with Azure DevOps
  - Ecosystem integrations for CI/CD and monitoring.

Azure – Data Science Virtual Machine (DSVM)

https://azure.microsoft.com/en-us/services/virtual-machines/data-science-virtual-machines/

- Pre-Configured virtual machines in the cloud for Data Science and AI Development
  - Python, R, ONNX, + Microsoft AutoML
  - https://docs.microsoft.com/en-us/azure/machine-learning/data-science-virtual-machine/tools-included
- Standard Configuration Standard_DS3_v2 – 4 vcups – 1 GiB
  - @ 328.12 USD/mo
- Expected to be used as a virtual desktop
- DSVM configuration
  - Demchmin :: BDIT****ndows





Site Reliability Engineering: Gluing together Data Analytics and Operation

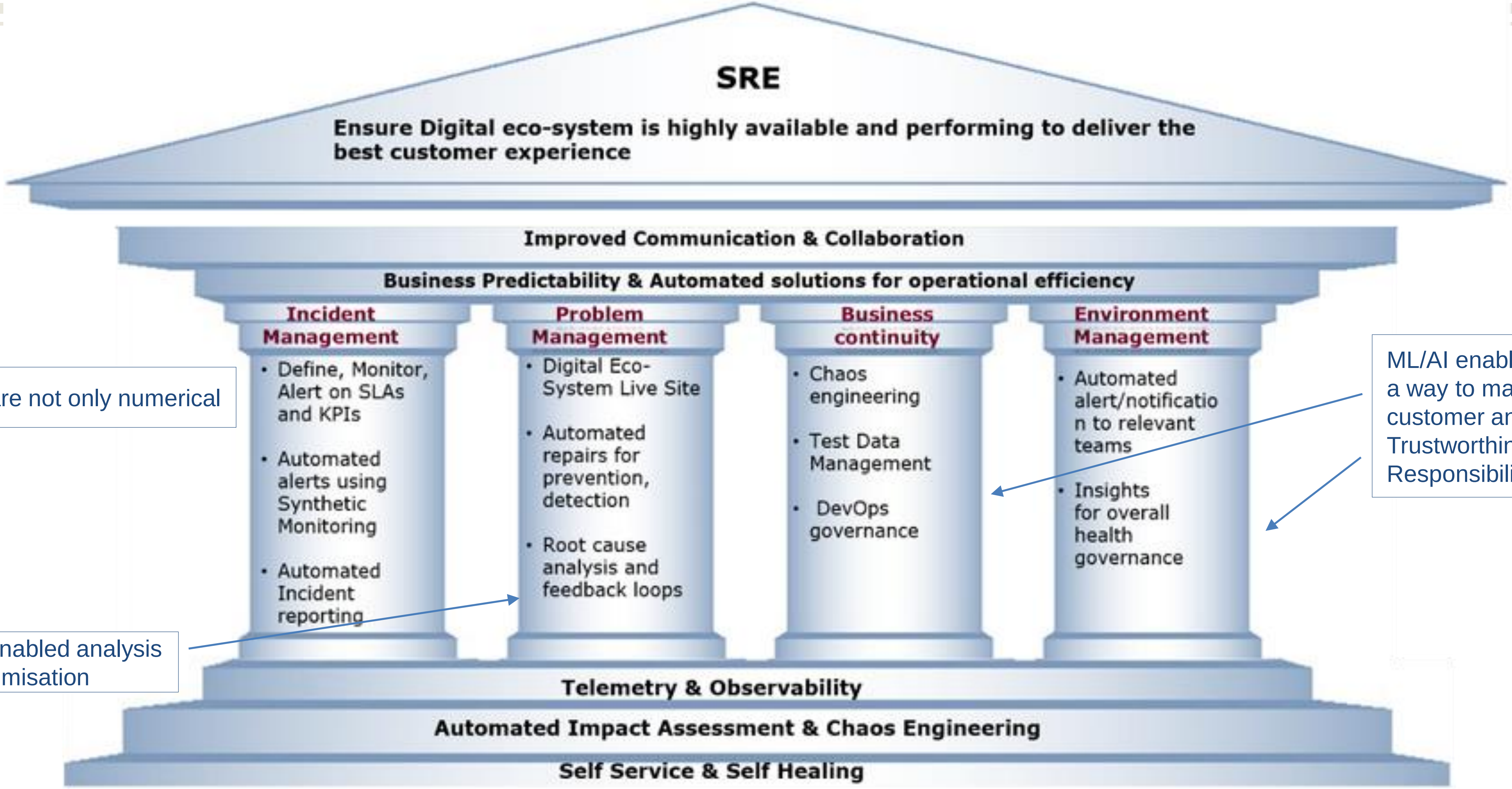
- Further shift to Operation stage in DevOps and service-oriented customer-facing approach
- Empowered by Data Analytics and DataOps
- To facilitate ML/AI operationalisation

SRE Pillars according to Capgemini [ref]

KPIs are not only numerical

ML/AI enabled monitoring as a way to manage complex customer and environment KPI: Trustworthiness, Sustainability and Responsibility

ML/AI enabled analysis and optimisation



SRE

Ensure Digital eco-system is highly available and performing to deliver the best customer experience

Improved Communication & Collaboration

Business Predictability & Automated solutions for operational efficiency

Incident Management

- Define, Monitor, Alert on SLAs and KPIs
- Automated alerts using Synthetic Monitoring
- Automated Incident reporting

Problem Management

- Digital Eco-System Live Site
- Automated repairs for prevention, detection
- Root cause analysis and feedback loops

Business continuity

- Chaos engineering
- Test Data Management
- DevOps governance

Environment Management

- Automated alert/notification to relevant teams
- Insights for overall health governance

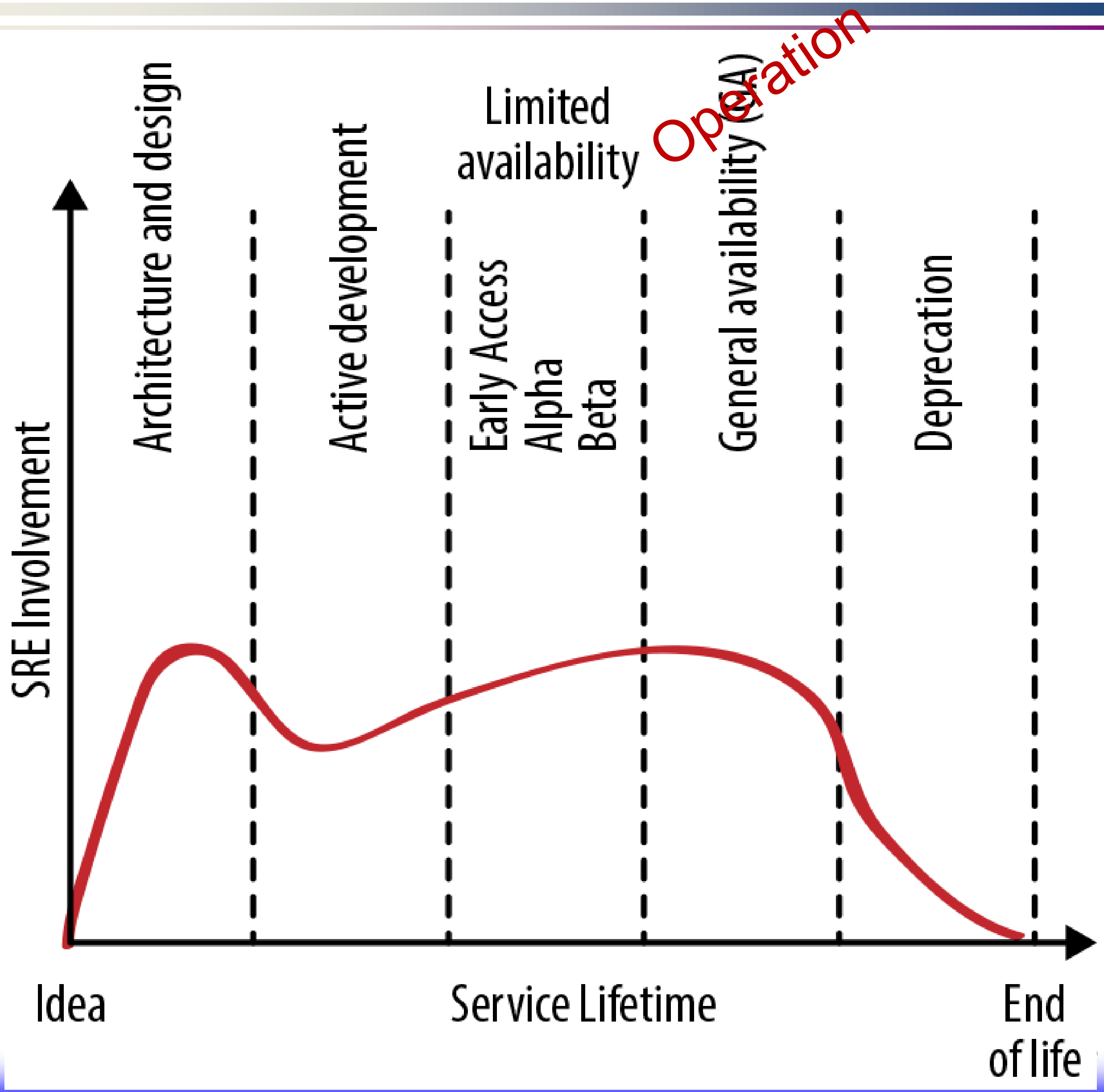
Telemetry & Observability

Automated Impact Assessment & Chaos Engineering

Self Service & Self Healing

[ref] https://www.capgemini.com/2020/08/site-reliability-engineering-2/

SRE and Data Analytics Engagement



What are benefits of such fusion/integration

- Continuous model improvement
  - Essential for operational AI and Digital Twins
- Monitoring and optimization on both measurable and assessable/ranked KPI

[ref] https://sre.google/workbook/engagement-model/

Discussion and Questions

- Recent trends to Data Analytics and ML Operationalisation
- Facilitated by digitalization and easy data concentration
- Role of and dependence on Big Data platform (cloud based)

From DataOps, DevOps to ModelOps

https://www.techzine.nl/blogs/data/432735/na-dataops-devops-nu-ook-modelops/

- What is DataOps by IBM - https://www.ibm.com/nl-en/analytics/dataops
- DataOps: Industrializing Data and Analytics Strategies for Streamlining the Delivery of Insights https://s3.amazonaws.com/eckerson/content_assets/assets/000/000/195/original/DataOPS.pdf?1534882627
- DataOps: Simplify Managing Complex Data Environments https://www.sentryone.com/dataops-overview
- DataOps is NOT Just DevOps for Data https://medium.com/data-ops/dataops-is-not-just-devops-for-data-6e03083157b7