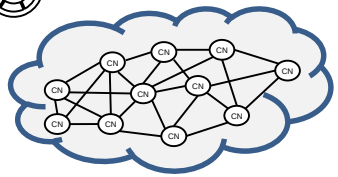
Cloud based solutions for Big Data:

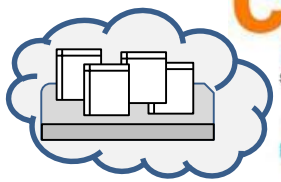
Cloud based Big Data Infrastructure and Platforms, SQL and NoSQL databases,

DevOps and DataOps

Yuri Demchenko

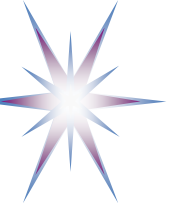
System and Networking Lab, UvA





Outline

- Definitions
  - Big Data, Cloud Computing, Data Science
- Standardisation
  - NIST, CSA, DMTF, IDS, RDA, DAMA. IEEE
- Big Data platforms and tools
  - Big Data Storage, SQL and NoSQL, modern databases
  - Apache Hadoop Ecosystem
  - AWS, Google Cloud Platform, Azure
- DevOps and DataOps
  - Cloud automation and monitoring tools
  - Azure DevOps services and tools
- Data Markets and Data Exchange
  - Data properties as economic goods
- Essential Big Data and Data Science Competences and Skills
  - EDISON Data Science Framework (EDSF)
- Learning and development resources
  - How to become a Data Scientist
- Discussion



Yuri Demchenko, Senior Researcher, Lecturer, UvA

- Graduated and PhD from National Technical University of Ukraine “Kiev Polytechnic Institute”
  - University of Amsterdam – since 2003
- Research areas
  - Big Data Infrastructure and Data Science platforms
  - Cloud architecture, cloud automation and DevOps
  - Cloud security and compliance
- Teaching courses (on campus and online)
  - Big Data Infrastructure and Technologies
  - Cloud powered Software Engineering and DevOps
  - Data Science Foundations, Professional Issues in Data Science
  - Security Engineering
- Recent projects
  - EDISON: Building the Data Science Profession for Europe
  - MATES: Digitalisation of the European Blue Economy
  - CYCLONE: Multi-cloud automation platform for cloud based applications
  - GEANT4 Research: Cloud aware networking infrastructure provisioning on-demand

Multiple aspects of Big Data

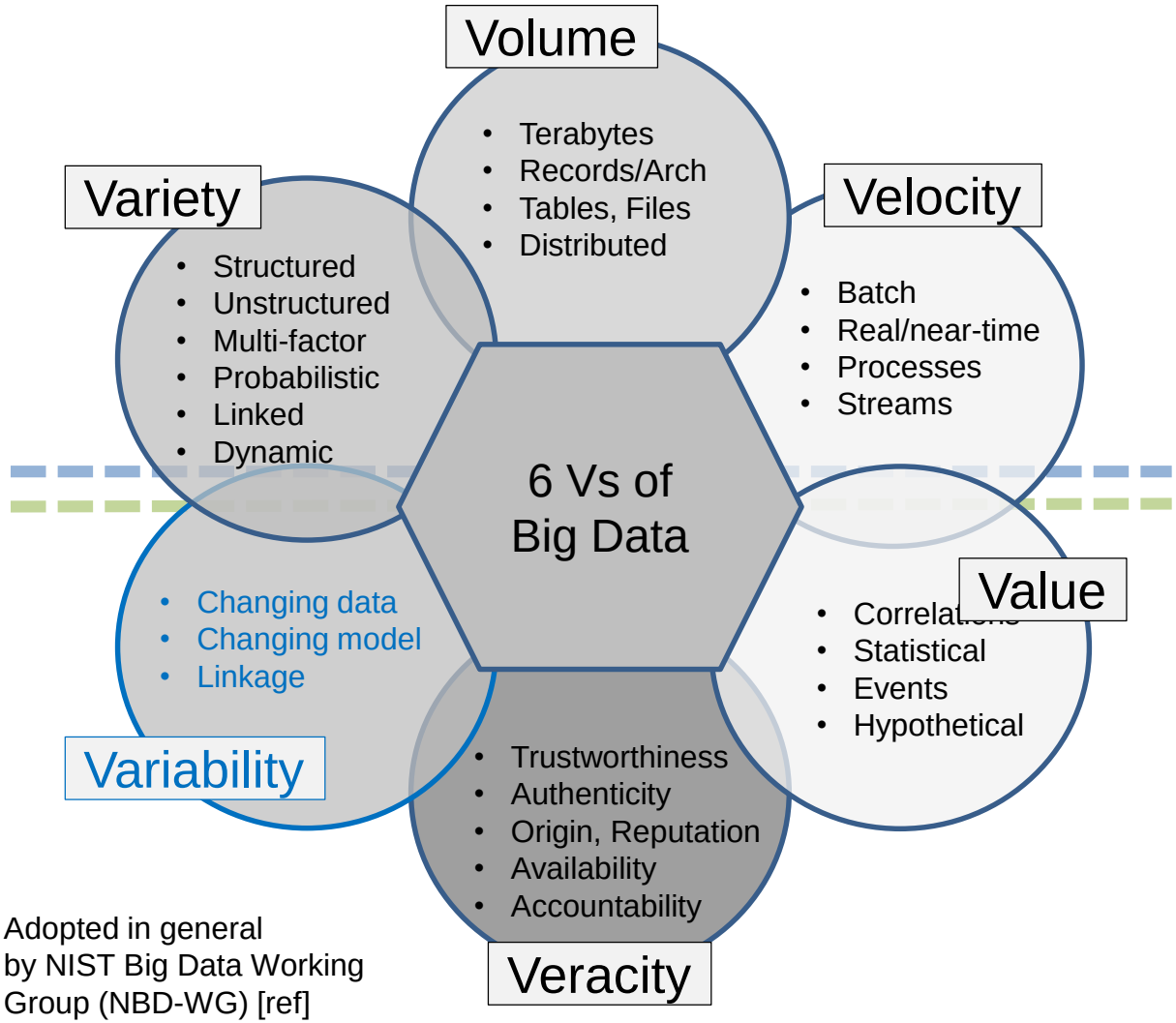
Big Data is a complex of technologies to enable handling of Big Data (storage, processing, transfer, security)

Big Data and multiple sources of data

- Social Media
- IoT
- Internet
- Science
- Industrial data
- Communication, voice

Data analytics blending with open and social media data

Big Data Properties: 6 (3+3) V's of Big Data



Adopted in general by NIST Big Data Working Group (NBD-WG) [ref]

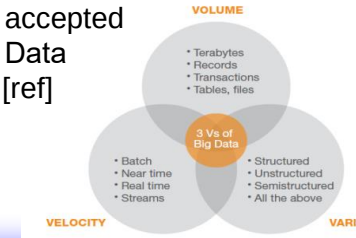
Generic Big Data Properties

- Volume
- Variety
- Velocity

Acquired Properties (after entering system)

- Value
- Veracity
- Variabilityfc AD

Commonly accepted 3V's of Big Data by Gartner [ref]



Big Data Definition: More than just BD properties

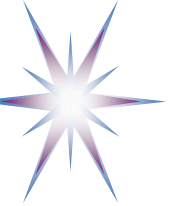
(1) Big Data Properties: 6V
- Generic: Volume, Variety, Velocity
- Ecquired: Value, Veracity, Variability

(2) New Data Models
- NoSQL
- Data linking, provenance and referral integrity
- Data Lifecycle and Variability/Evolution

(3) New Analytics
- Real-time/streaming analytics, interactive and machine learning analytics

(4) New Infrastructure and Tools
- High performance Computing, Storage, Network
- New Data Centric (multi-stakeholder) service models
- New Data Centric security models for trusted infrastructure and data processing

(5) Source and Target
- High velocity/speed data capture from variety of sensors and data sources
- Data delivery to different visualisation and actionable systems and consumers
- Fully digitised input and output, (ubiquitous) sensor networks, full digital control



Cloud Computing, Big Data, Data Science

- Cloud Computing
  - Infrastructure/Platform/Software as a Service (IaaS/PaaS/SaaS)
  - Private, public, hybrid, community, federated

- Big Data
  - Technology domain to enable handling of Big Data (storage, processing, transfer, security)
  - Big Data properties and new data centric models

- Data Science (NIST)
  - **Data science** is the extraction of actionable knowledge directly from data through a process of discovery, or hypothesis formulation and hypothesis testing.
  - Data science combines concepts and methods from multiple disciplines to enable whole data lifecycle to bring value to business

Technology Maturity and Standardisation

- NIST: Cloud Computing (2008-2013) and Big Data (2013-now)
  - Cloud Computing Reference Architecture (CCRA)
  - Big Data Reference Architecture (BDRA)
- DMTF – Distributed Management Task Force
  - Cloud Information Model (CIM)
  - Open Virtualisation Format (OVF)
- CSA – Cloud Security Alliance
  - Cloud Compliance and Big Data Security Controls
- RDA – Research Data Alliance
  - PID, Data Factories, Data Registries
- DAMA – Data Management Association
  - Data Management Body of Knowledge (DMBOK)
- Industrial Data Space Association (IDS)
  - Industrial Data Space Architecture

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

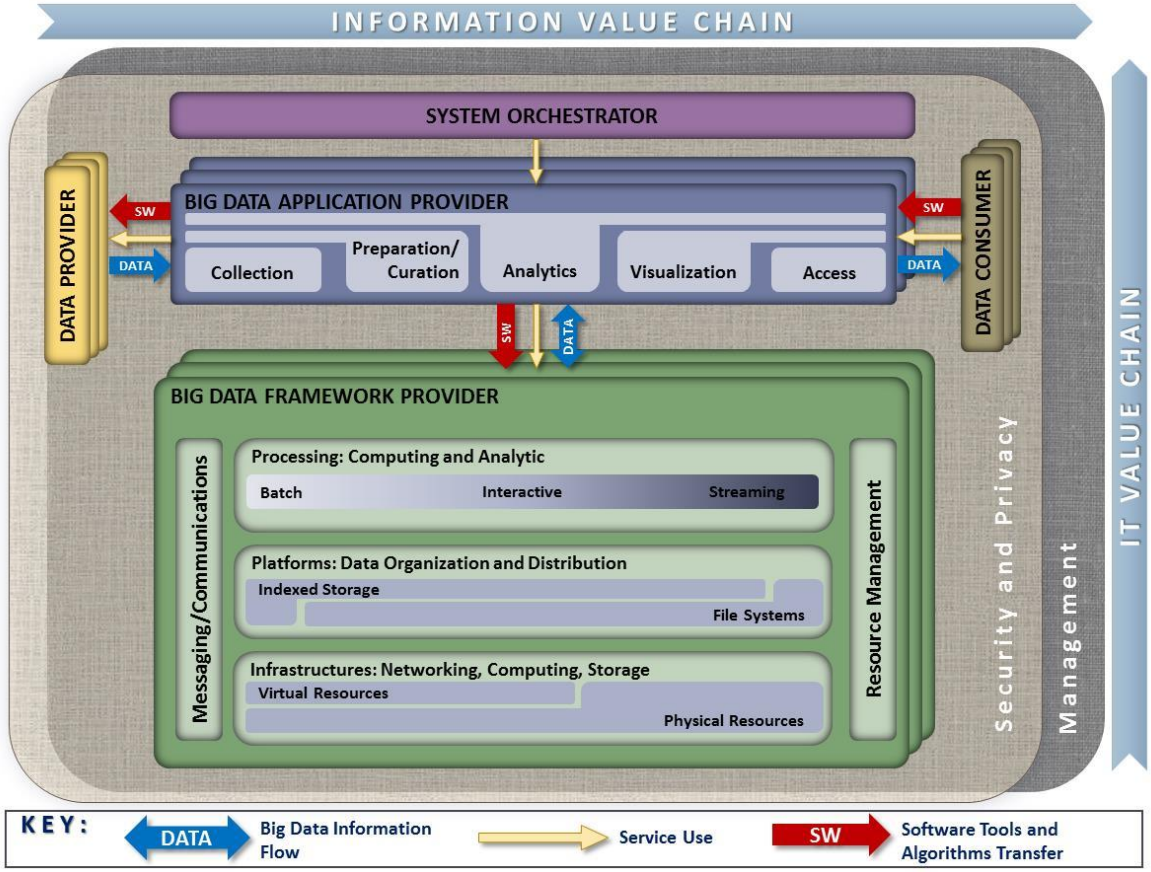
- NIST Big Data Working Group (NBD-WG) is leading the development of the Big Data Technology Roadmap - http://bigdatawg.nist.gov/home.php
  - Built on experience of developing the Cloud Computing standards fully accepted by industry
- Set of documents delivered September 2014 (to be published as NIST Special Publication documents) - http://bigdatawg.nist.gov/V1_output_docs.php

  Volume 1: NIST Big Data Definitions
  Volume 2: NIST Big Data Taxonomies
  Volume 3: NIST Big Data Use Case & Requirements
  Volume 4: NIST Big Data Security and Privacy Requirements
  Volume 5: NIST Big Data Architectures White Paper Survey
  Volume 6: NIST Big Data Reference Architecture
  Volume 7: NIST Big Data Technology Roadmap

- 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

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.

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

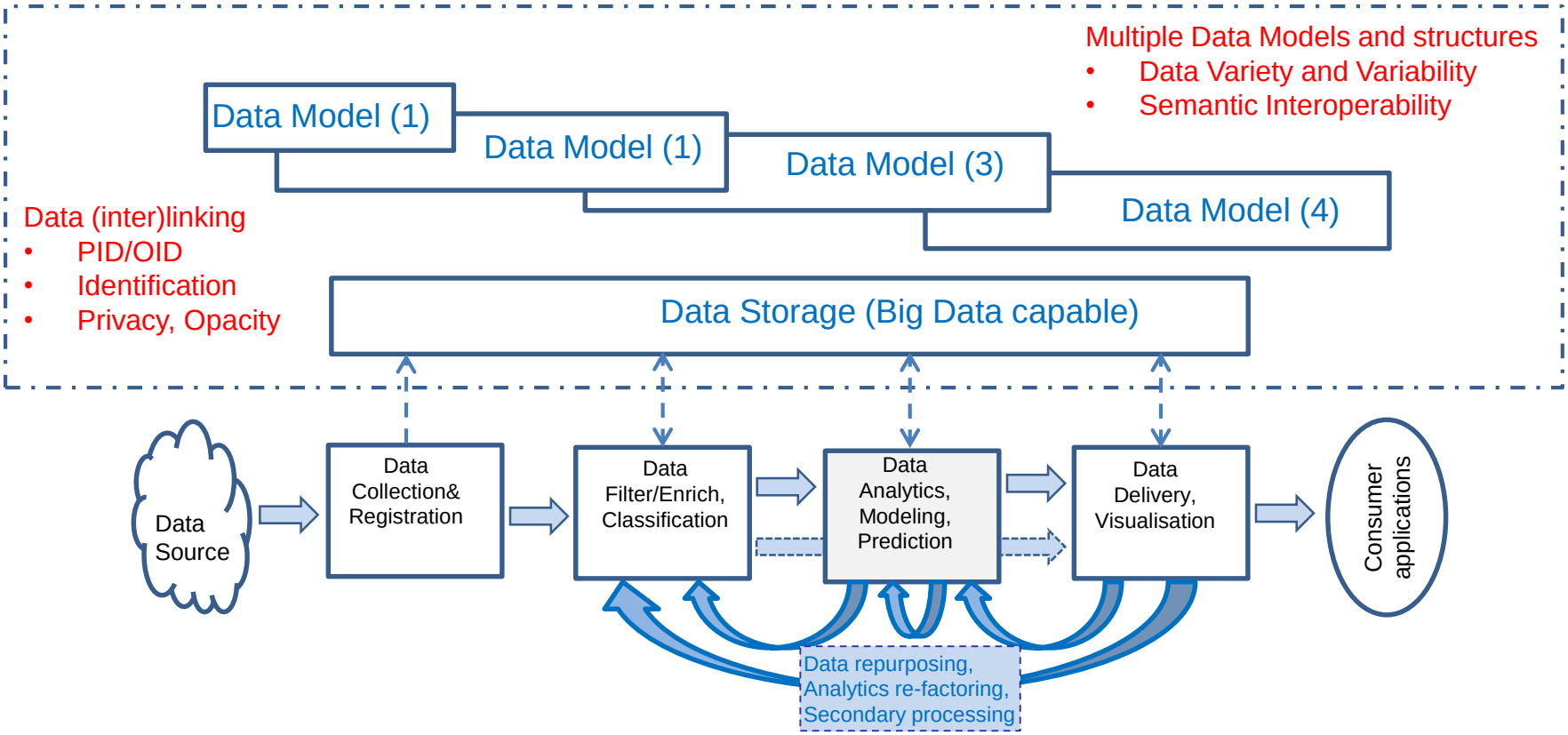
Big Data Lifecycle and Applications Provider activities

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

Big Data Ecosystem includes all components that are involved into Big Data production, processing, delivery, and consuming

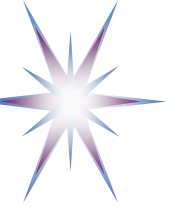
[ref] Volume 6: NIST Big Data Reference Architecture. http://bigdatawg.nist.gov/V1_output_docs.php

Data Lifecycle/Transformation Model



- Data Model changes along data lifecycle or evolution
- Data provenance is a discipline to track all data transformations along lifecycle

- Identifying and linking data
  - Persistent data/object identifiers (PID/OID)
  - Traceability vs Opacity
  - Referral integrity



SQL and NoSQL

NoSQL definition (www.nosql-database.org):

- Next Generation Databases mostly addressing some of the points: being **non-relational, distributed, open-source** and **horizontal scalable**.
- Other characteristics apply: **schema-free, easy replication support, simple API, eventually consistent / BASE** (not ACID), a **huge data amount**, others

- ACID/SQL vs BASE/NoSQL
  - ACID Semantics: **A**tomic, **C**onsistent, **I**solated, **D**urable
  - BASE Semantics: **B**asically **A**vailable - **S**oft State - **E**ventual Consistency

NoSQL Distinguishing Characteristics

- Large data volumes
  - Web scale “Big Data”
- Scalable replication and distribution
  - Potentially thousands of machines
  - Potentially distributed around the world
- Queries need to return answers quickly
  - Not necessary precisely
  - Employing probabilistic search/decision
- Mostly query, few updates
- Asynchronous Inserts and Updates
- Schema - free
- Paradigm shift from ACID transaction properties to BASE
- CAP Theorem – NoSQL database types
- Open source development

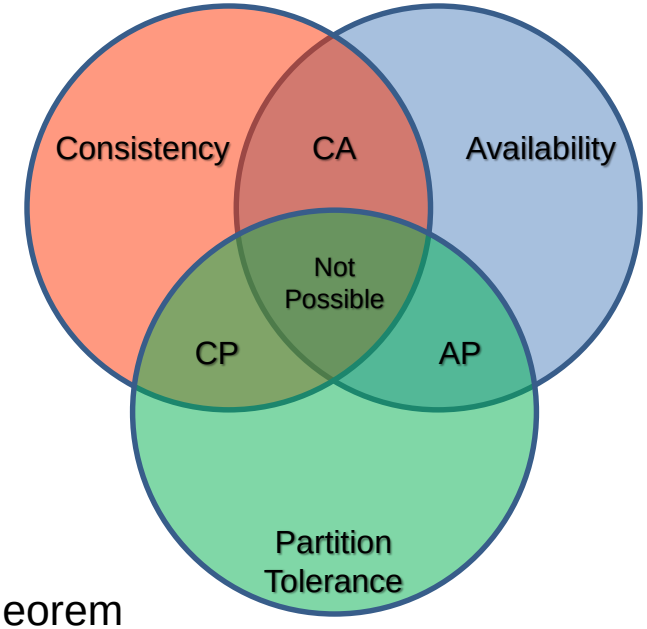
Brewer's CAP Theorem

Brewer's (CAP) Theorem (original formulation) [ref]

> *"There are three core systemic requirements that exist in a special relationship when it comes to designing and deploying applications in a distributed environment."*

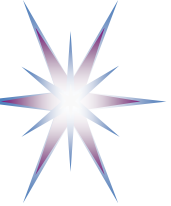
A distributed system can support only two of the following characteristics:

- Consistency
  - All nodes see the same data at the same time
- Availability
  - Node failures do not prevent survivors from continuing to operate
- Partition tolerance
  - The system continues to operate despite arbitrary message loss



[ref] http://www.julianbrowne.com/article/viewer/brewers-cap-theorem

The CAP Theorem Dilemma in Cloud Database and Storage



Main NoSQL Database Types and Existing Implementations

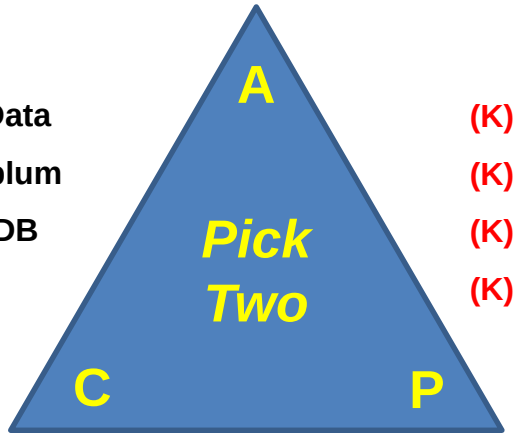
- **Column Store**: Each storage block contains data from only one column
  - BigTable by Google, Apache HBase, Cassandra
  - Work natively with HDFS and Hadoop
- **Document Store**: Store documents (in particular XML) made up of tagged elements
  - MongoDB, CouchDB, RaptorDB, CosmosDB
- **Key-Value Store**: Hash table of keys with arbitrary data as content/value
  - Memcached, Membase, Accumulo, Amazon DynamoDB
- **Graph Databases**: Store data in graph data in Triplestores or Quadstores
  - Neo4J, FlockDB, GraphDB

Visualizing the CAP Theorem and it's members

Configurable consistency:
Azure CosmosDB (K,C,D,G), AWS Aurora (R), Google Spanner (R)

Data Models:
(R) Relational (Comparison)
(K) Key Value
(C) Column-Oriented/Tabular
(D) Document-Oriented
(G) Graph

Availability
Each client can always read and write



CA

(R) RDBMs such as MySQL, Oracle, DB2, PostgreSQL, SQL Server, etc.
(R) Aster Data
(R) Greenplum
(D) CouchDB
(C) Vertica

AP

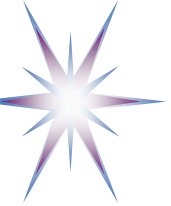
(K) Dynamo
(K) Voldemort
(K) Tokyo Cabinet
(K) KAI
(C) Cassandra
(D) SimpleDB
(D) CouchDB
(D) Riak

Consistency
All clients always have the same view of data

Partition Tolerance
The system works well despite physical network partitions

CP

(C) BigTable
(C) Hypertable
(C) Hbase
(D) MongoDB
(D) Terastore
(K) Scalaris
(K) BerkeleyDB
(K) MemcacheDB
(K) Redis
(R) VoltDB



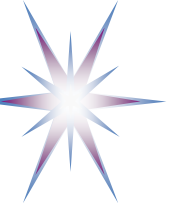
Modern Cloud Databases and CAP Theorem Challenges

- Google Spanner SQL database
- AWS Aurora Big SQL database
- Azure CosmosDB (NoSQL multi-data format database with underlying blob storage and HDFS)

Cloud Spanner - Mission Critical RDBMS

- SQL Database
- Best of both worlds
- Strong consistency and high availability worldwide

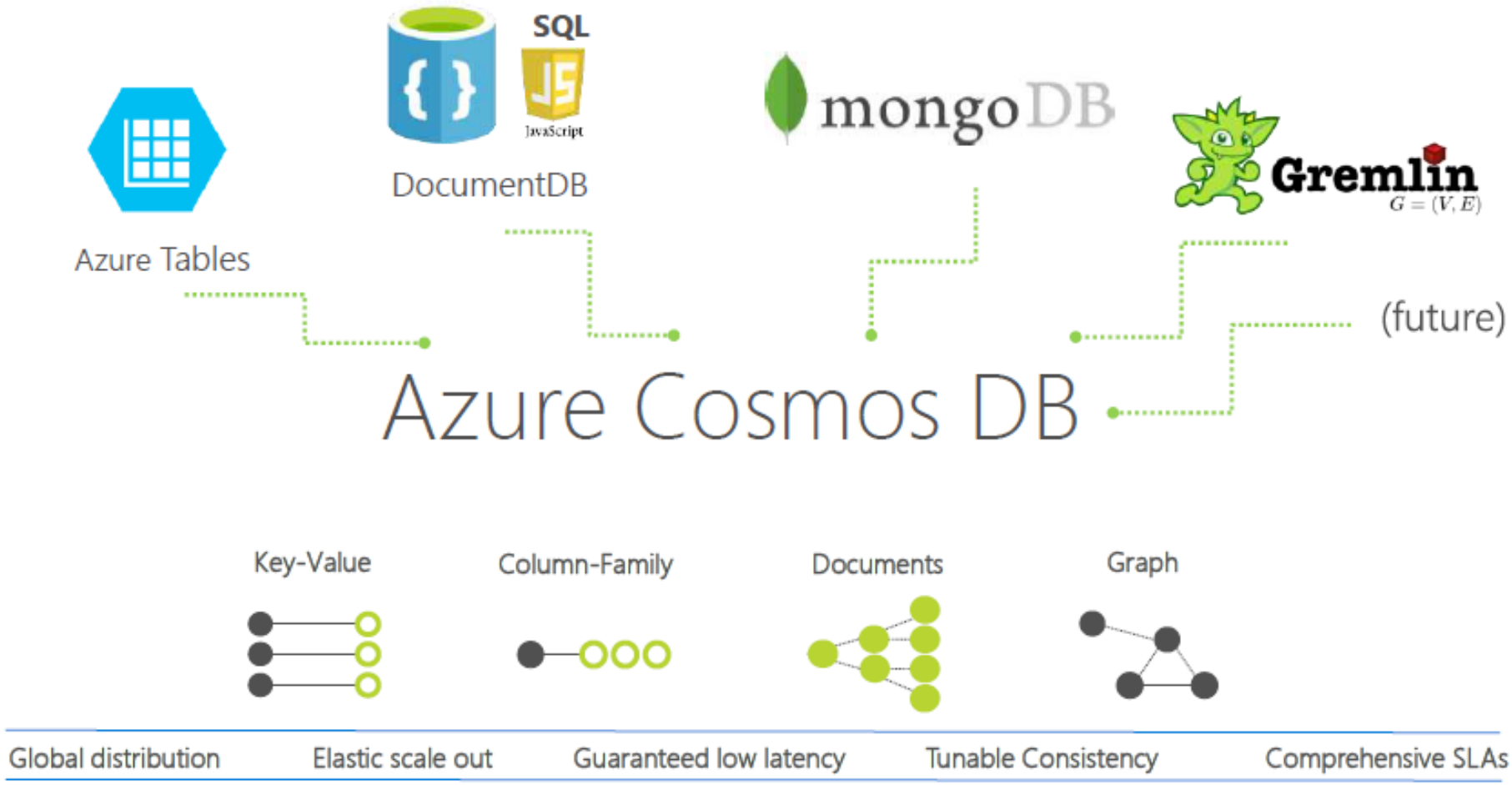
	CLOUD SPANNER	TRADITIONAL RELATIONAL	TRADITIONAL NON-RELATIONAL
Schema	✓ **Yes**	✓ Yes	✗ No
SQL	✓ **Yes**	✓ Yes	✗ No
Consistency	✓ **Strong**	✓ Strong	✗ Eventual
Availability	✓ **High**	✗ Failover	✓ High
Scalability	✓ **Horizontal**	✗ Vertical	✓ Horizontal
Replication	✓ **Automatic**	↻ Configurable	↻ Configurable



Amazon Aurora – Big SQL database

- High performance and scalability
- High availability and durability
- High security
- **MySQL and PostgreSQL** Compatible
- Fully managed
- Migration support

Azure CosmosDB (former DocumentDB) – Multi-model global distributed database



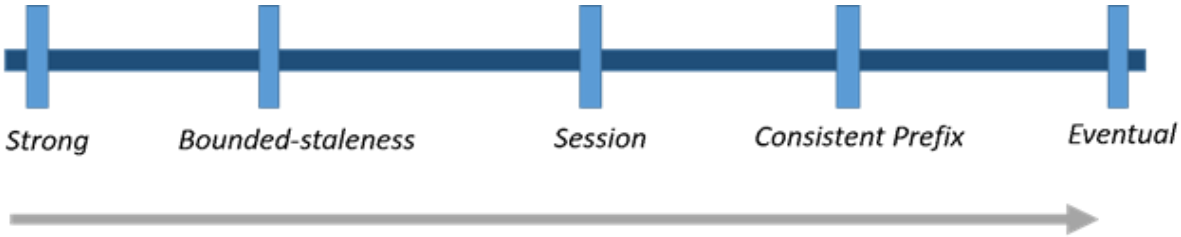
CosmosDB: Consistency and latency

Five Consistency Models

- Helps navigate Brewer's CAP theorem
- Intuitive Programming
  - Tunable well-defined consistency levels
  - Override on per-request basis
- Clear PACELC tradeoffs
  - Partition – Availability vs Consistency
  - Else – Latency vs Consistency

- Guaranteed low latency at P50 and P99 percentile
- Globally distributed with requests served from local region
- Write optimized, latch-free database
- Automatic Indexing

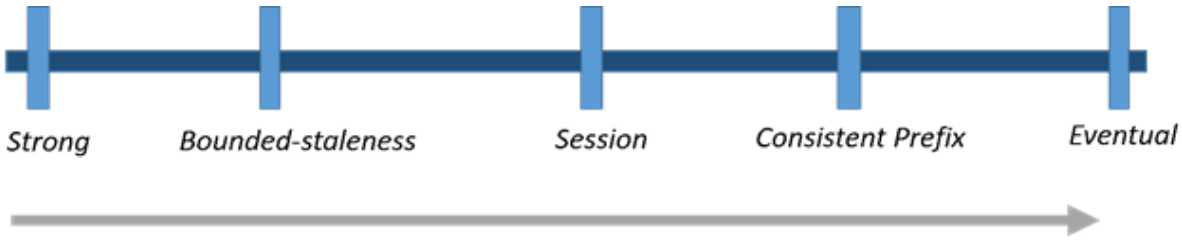
	Reads (1KB)	Indexed Writes (1KB)
P50	<2ms	<6ms
P99	<10ms	<15ms

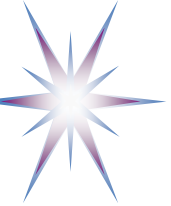


CosmosDB: Consistency levels and guarantees

Consistency Level	Guarantees
Strong	Linearizability
Bounded Staleness	Consistent Prefix. Reads lag behind writes by k prefixes or t interval
Session	Consistent Prefix. Monotonic reads, monotonic writes, read-your-writes, write-follows-reads
Consistent Prefix	Updates returned are some prefix of all the updates, with no gaps
Eventual	Out of order reads

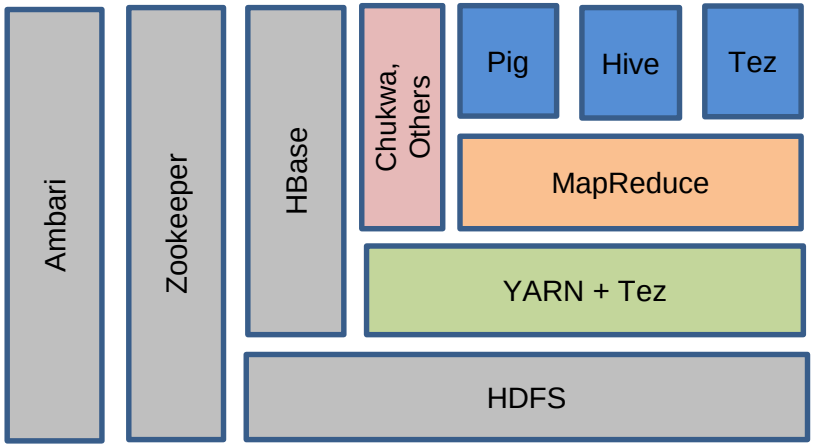
Lower latency, higher availability, better read scalability





Hadoop and Big Data processing Platforms

Apache Hadoop (Release 2.7+)



Apache Hadoop software stack includes the following main modules:

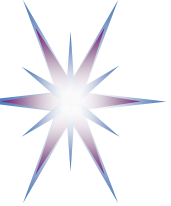
- **Hadoop Common**: The common utilities that support the other Hadoop modules and includes utilities and drivers to support different computer cluster and language platforms.
- **HDFS**: Hadoop Distributed File System optimized for large scale storage and processing of data on commodity hardware
- **Hadoop YARN:** A framework for job scheduling and cluster resource management.
- **Hadoop MapReduce:** A YARN-based system for parallel processing of large data sets.

Other Hadoop-related projects at Apache include:

- **HBase:** A distributed column oriented database that supports structured data storage for large tables
- **Hive:** A data warehouse system that provides data aggregation and querying.
- **Pig:** A high-level data-flow language and execution framework for parallel computation.
- **Mahout:** A scalable machine learning and data mining library.
- **Tez:** A generalized data-flow programming framework, built on Hadoop YARN, which provides a powerful and flexible engine to execute an arbitrary DAG of tasks to process data for both batch and interactive use-cases.
- **ZooKeeper**: A scalable coordination service for distributed applications.
- **Spark:** A fast and general compute engine for Hadoop data. Spark provides a simple and expressive programming model that supports a wide range of applications, including ETL, machine learning, stream processing, and graph computation.
- **Avro:** A data serialization system that supports rich data structures
- **Solr:** Data visualisation tools, dashboard design.
- **Hue:** User GUI
- **Ambari:** A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters
- **Flink, Storm, Flume, Sqoop, Oozie**

Hive and Pig Latin

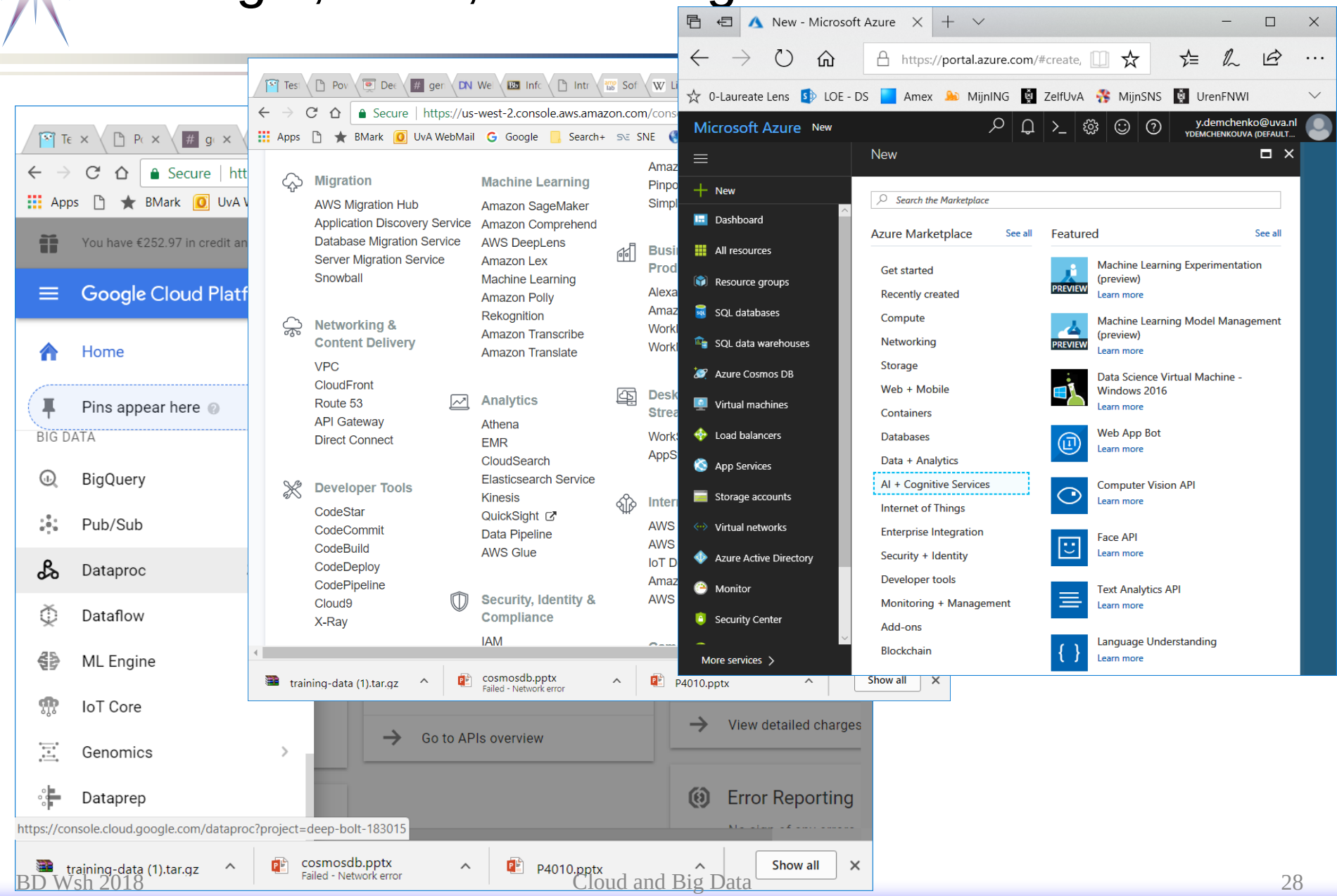
- Solution: Provide higher-level data processing languages
- Hive: Data warehousing application in Hadoc
  - Query language is HQL, variant of SQL
  - Tables stored on HDFS as flat files
  - Developed by Facebook, now open source
- Pig: Large-scale data processing system
  - Scripts are written in Pig Latin, a dataflow language
  - Developed by Yahoo!, now open source
  - Roughly 1/3 of all Yahoo! internal jobs
- Purpose:
  - Provide higher-level language to facilitate large-data processing
  - Higher-level language “compiles down” to Hadoop jobs



Cloud based Big Data Platforms and Solutions

- Amazon Web Services (AWS)
- Google Cloud Platform (GCP)
- Microsoft Azure

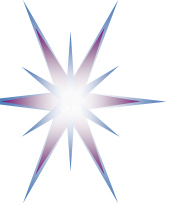
Google, AWS, Azure Big Data Stacks



AWS Cloud Big Data Services

AWS Cloud offers the following services and resources for Big Data processing:

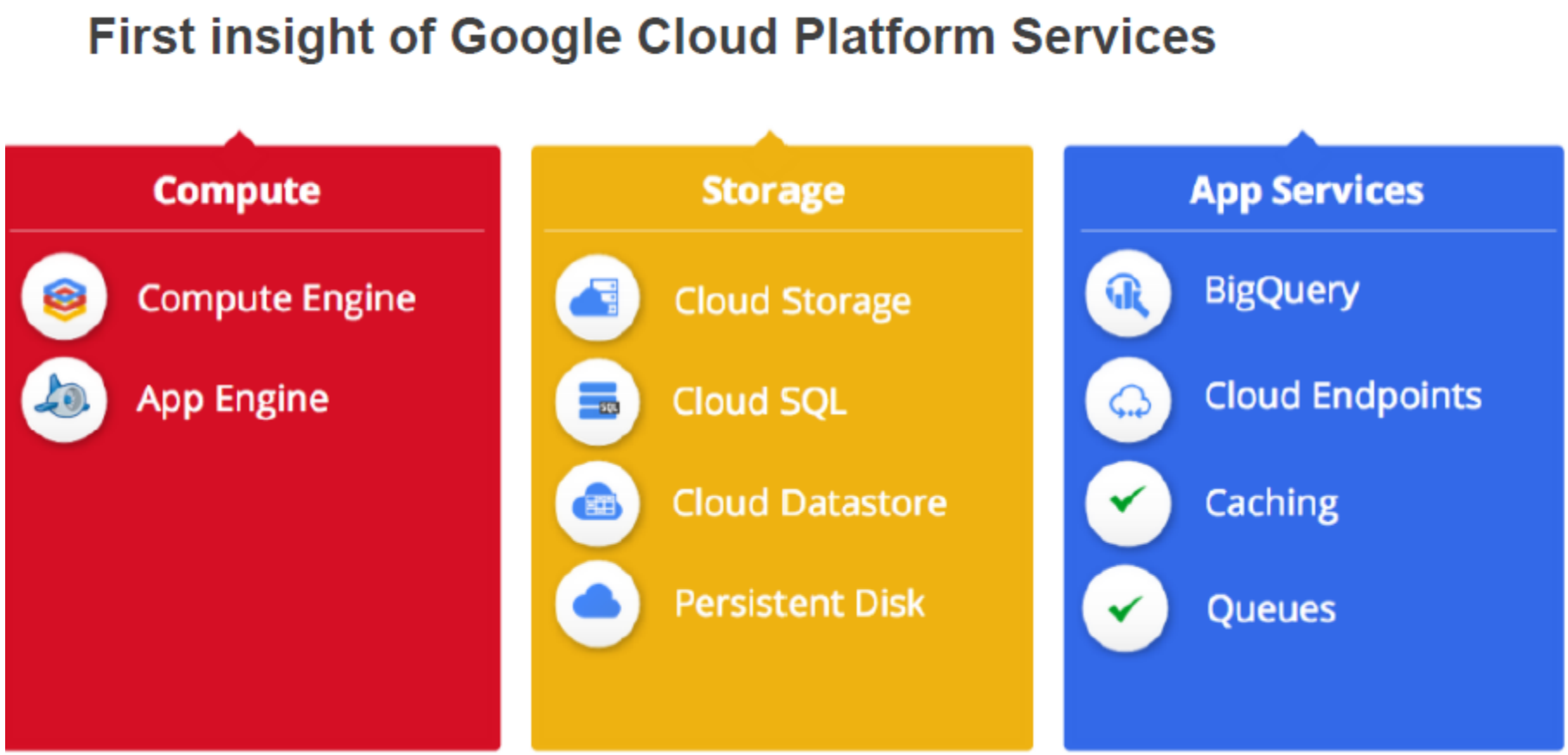
- EC2 Virtual Machine (VM) instances for HPC optimized for computing (with multiple cores) and with extended storage for large data processing.
- **Amazon Elastic MapReduce (EMR)** provides the Hadoop framework on Amazon EC2 and offers a wide range of Hadoop related tools.
- **Amazon Kinesis** is a managed service for real-time processing of streaming big data (throughput scaling from megabytes to gigabytes of data per second and from hundreds of thousands different sources).
- **Amazon DynamoDB** highly scalable NoSQL data stores with sub-millisecond response latency.
- **Amazon Aurora** scalable SQL/relational database.
- Amazon Redshift fully-managed petabyte-scale Data Warehouse in cloud at cost less than $1000 per terabyte per year.
- Amazon Machine Learning
  - Machine Learning (Artificial Intelligence) based services (Lex, Translate, Recognition, etc.)



Google Cloud Platform (GCP)

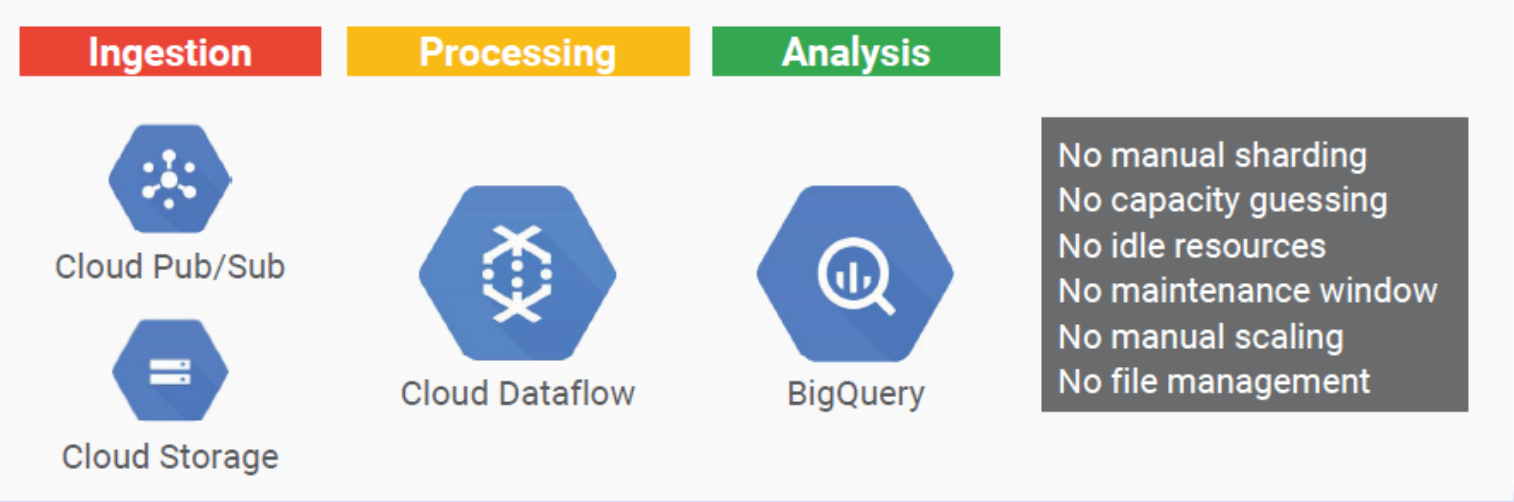
- Compute
  - AppEngine
  - Google Functions (serverless with Node.js)
- Storage: Static, sharing, backup, for applications and computation
  - Cloud Spanner SQL database
- Big Data
  - BigQuery – Hadoop Data Warehouse
- Machine Learning services
  - Translate
  - Prediction
- Cloud endpoints

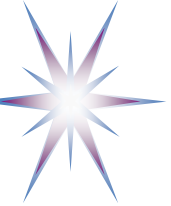
Google Cloud Platform (GCP) structure



Google Machine Learning

- Support all stages of ML workflow
- Key models exposed via APIs (Democratizing Machine Learning)
- Multiple Tensorflow ML models in use (portable)
- Dataprep: Serverless platform for all stages of the analytics data lifecycle



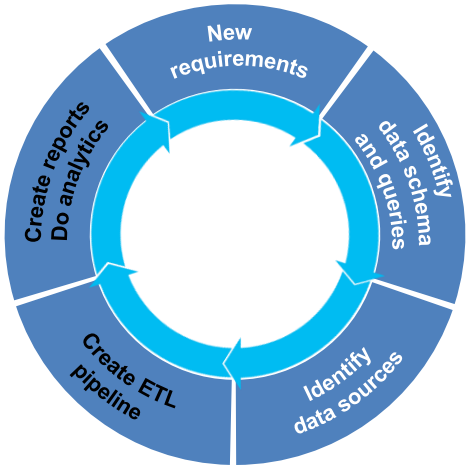


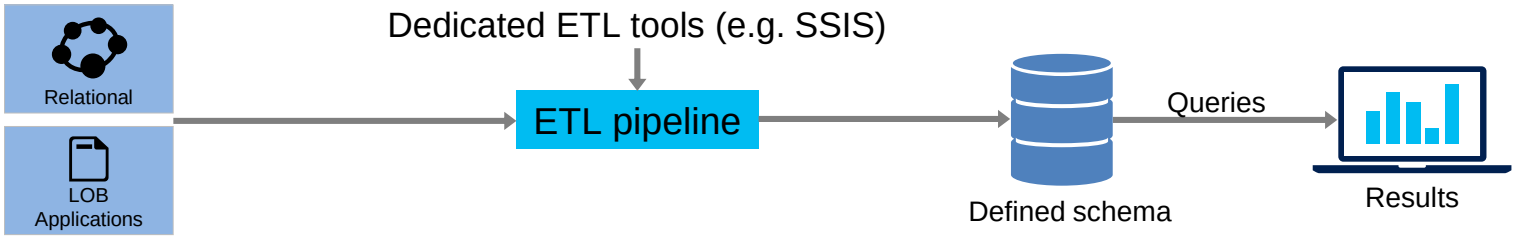
Microsoft Azure Big Data Services

- New Big Data data-centric thinking
- Azure Data Lakes
- HDInsight

Traditional business analytics process

1. Start with end-user requirements to identify desired reports and analysis
2. Define corresponding database schema and queries
3. Identify the required data sources
4. Create a Extract-Transform-Load (ETL) pipeline to extract **required data** (curation) and **transform it to target schema** (‘*schema-on-write*’)
5. Create reports, analyze data

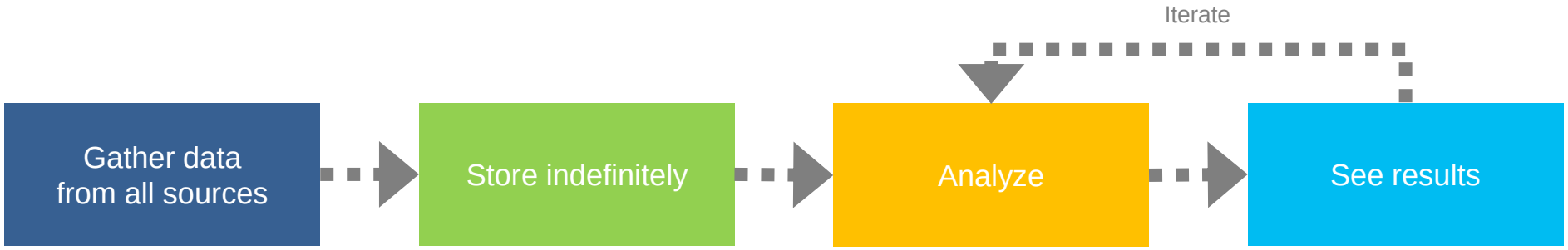




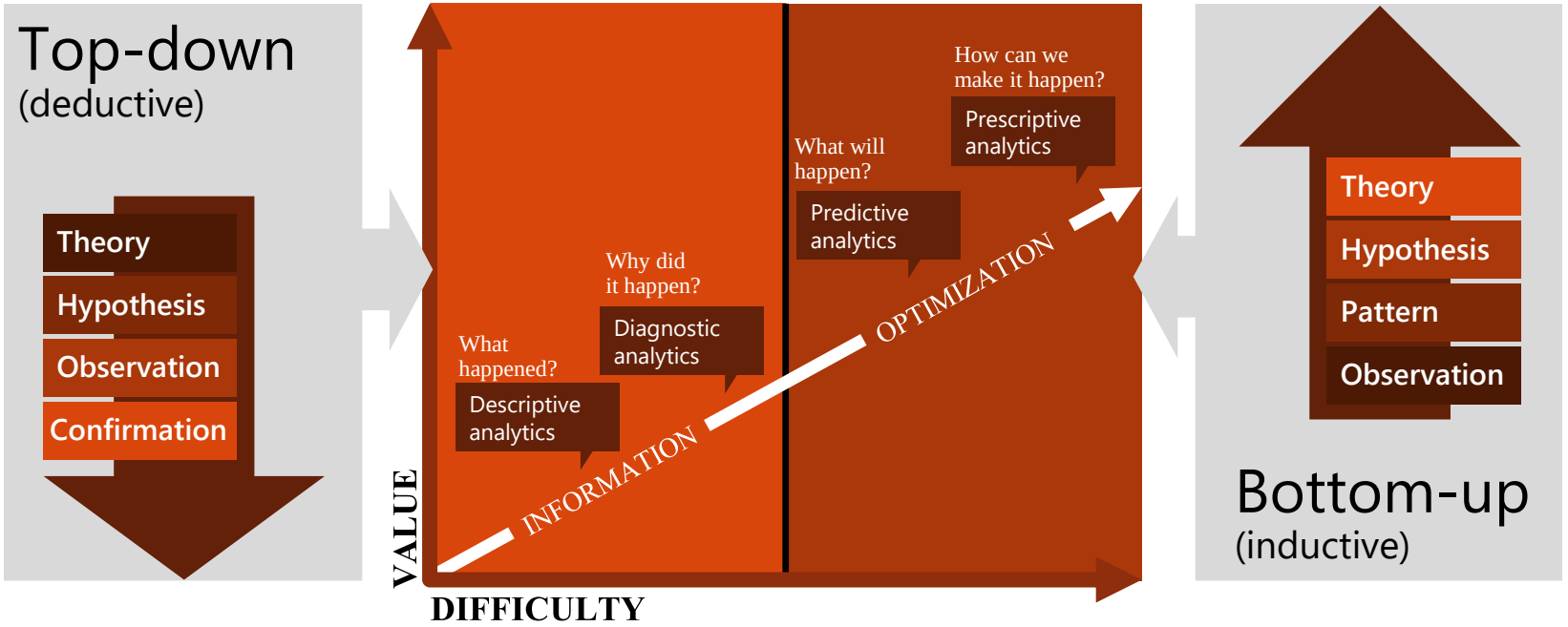
All data not immediately required is discarded or archived

New Big Data thinking: All data has value

- All data has potential value
- **Data hoarding** (data acquisition and reluctantly deleting)
- No defined schema—stored in native format
- Schema is imposed and transformations are done at query time *(schema-on-read).*
- Apps and users interpret the data as they see fit



Two approaches to information management for analytics: Top-down and bottom-up



- Data Warehousing uses a top-down approach
- The Data Lake uses a bottom-up approach
  - Ingest all data – Store all data – Do analysis (on purpose)

Data Lakes - Definition

Data Lake allows an organization to store all of their data, structured and unstructured, in one, centralized repository.

- Since data can be **stored as-is**, there is no need to convert it to a predefined schema and you no longer need to know what questions you want to ask of your data beforehand.

A Data Lake should support the following capabilities:

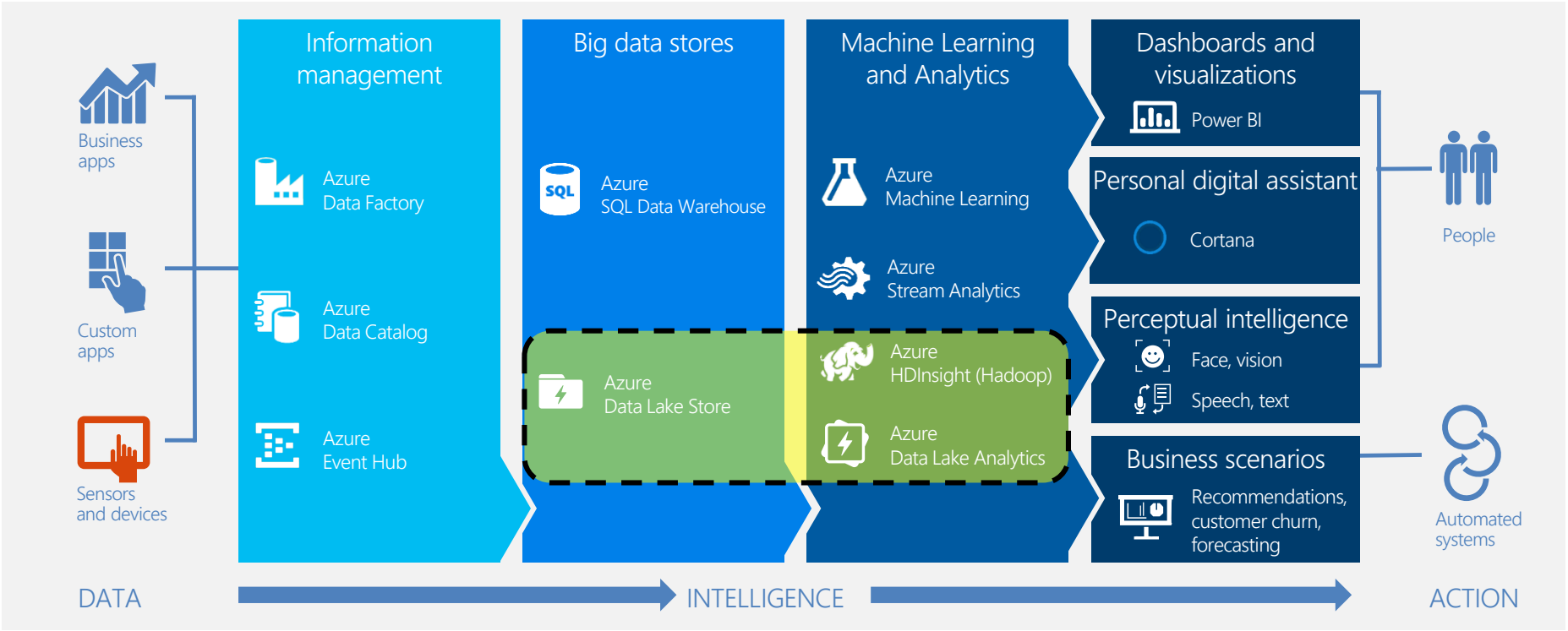
- Collecting and storing any type of data, at any scale and at low costs
- Securing and protecting all of data stored in the central repository
- Searching and finding the relevant data in the central repository
- Quickly and easily performing new types of data analysis on datasets
- Querying the data by defining the data's structure at the time of use (schema on read)

Data Lake layers

- **Raw data layer –** Raw events are stored for historical reference. Also called staging layer or landing area
- **Cleansed data layer –** Raw events are transformed (cleaned and mastered) into directly consumable data sets. Aim is to uniform the way files are stored in terms of encoding, format, data types and content (i.e. strings). Also called conformed layer
- **Application data layer –** Business logic is applied to the cleansed data to produce data ready to be consumed by applications (i.e. DW application, advanced analysis process, etc). Also called workspace layer or trusted layer

Still need data governance so your data lake does not turn into a data swamp!

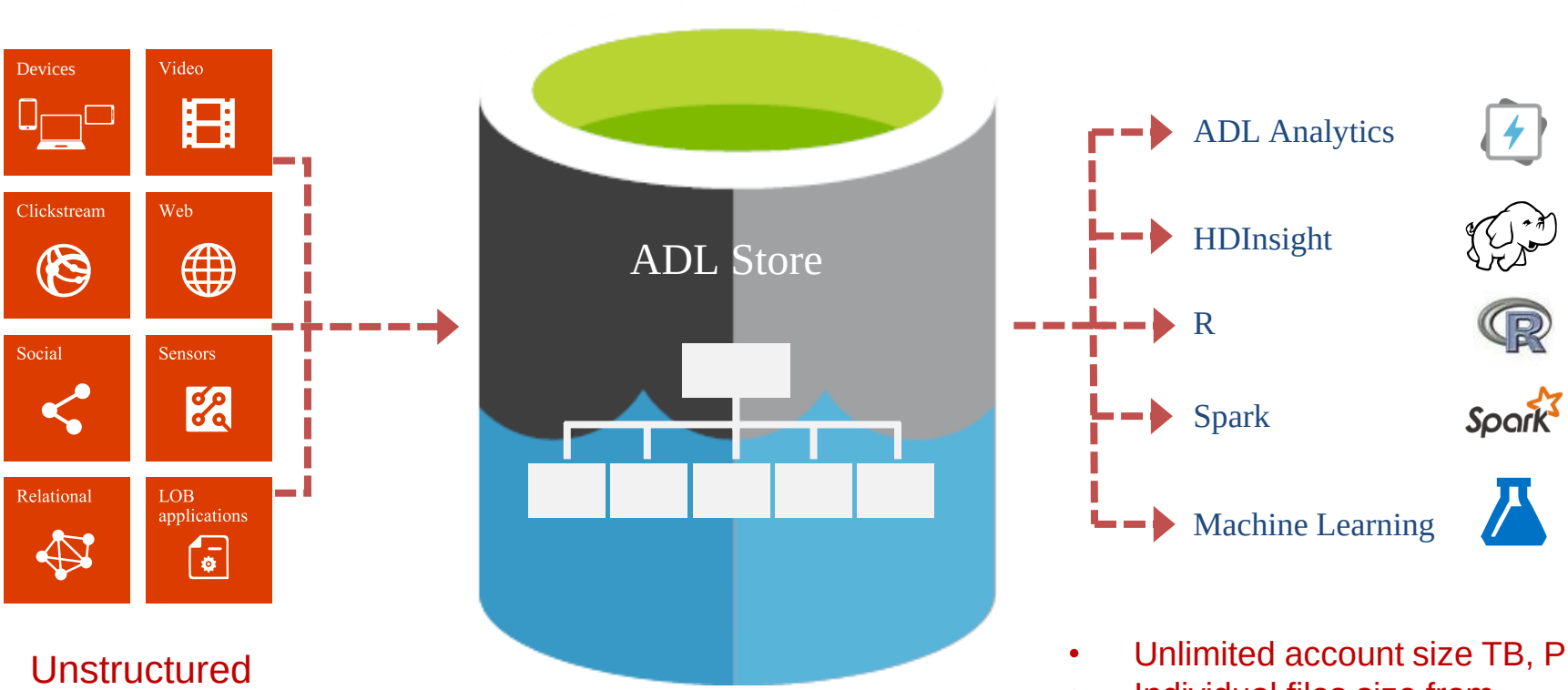
Azure Data Lake in Azure Big Data Stack



- Azure Data Lake Store (ADLS)
- Azure Data Lake Analytics (ADLA)
- Part of Cortana Analytics Suite

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

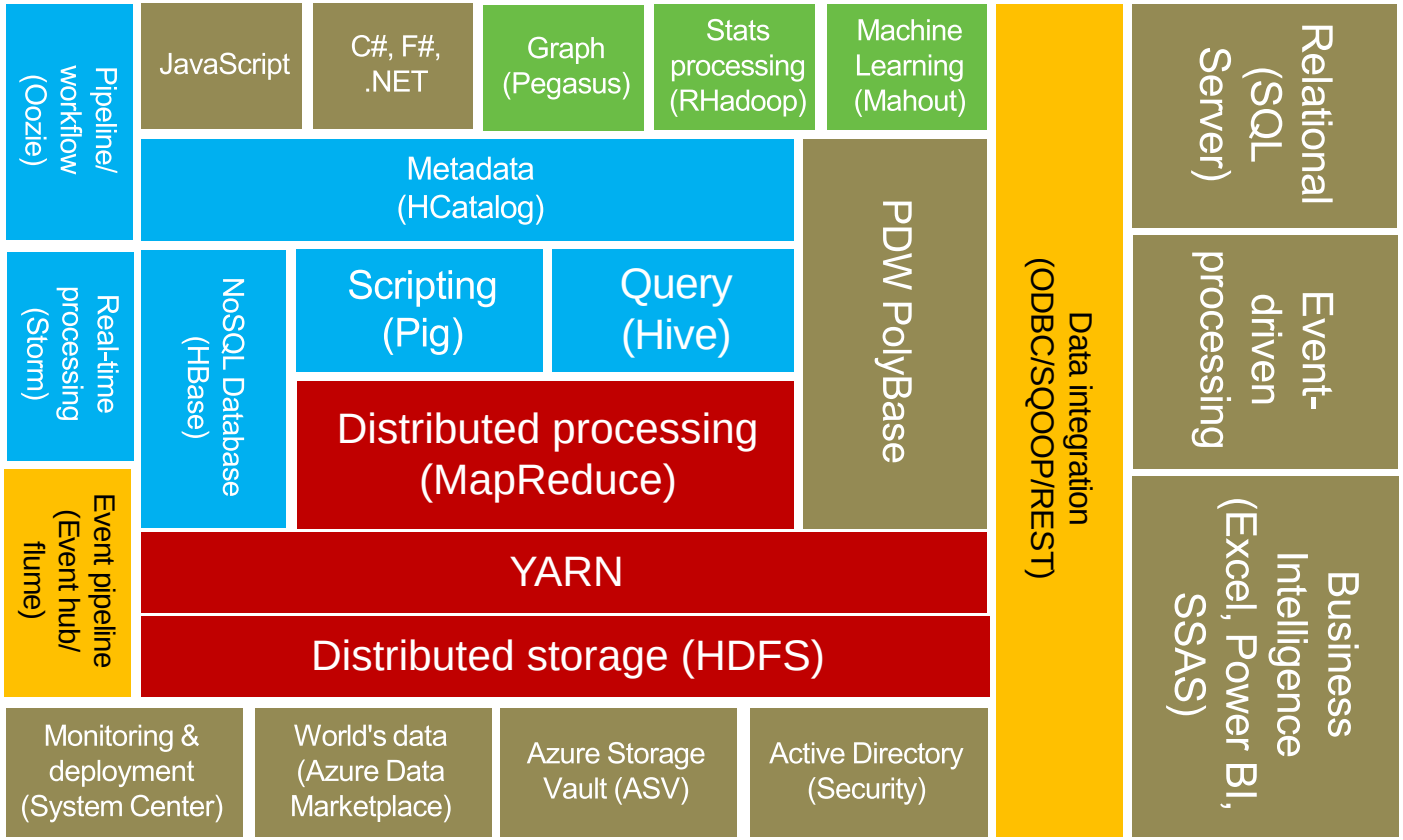
Azure HDInsight – What is it?

A standard Apache Hadoop distribution offered as a managed service on Microsoft Azure

- Based on Hortonworks Data Platform (HDP)
- Provisioned as clusters on Azure that can run on Windows or Linux servers
- Offers capacity-on-demand, pay-as-you-go pricing model
- Integrates with:
  - Azure Blob Storage and Azure Data Lake Store for Hadoop File System (HDFS)
  - Azure Portal for management and administration
  - Visual Studio for application development tooling

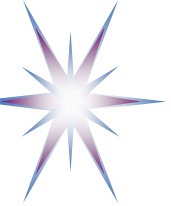
In addition to the core, HDInsight supports the Hadoop ecosystem

HDInsight and Hadoop ecosystem (2017)



Legend
Red = Core Hadoop
Blue = Data processing
Gray = Microsoft integration points and value adds
Orange = Data movement
Green = Packages

HDInsight supports Mahout, HBase, Storm, Hive, Spark



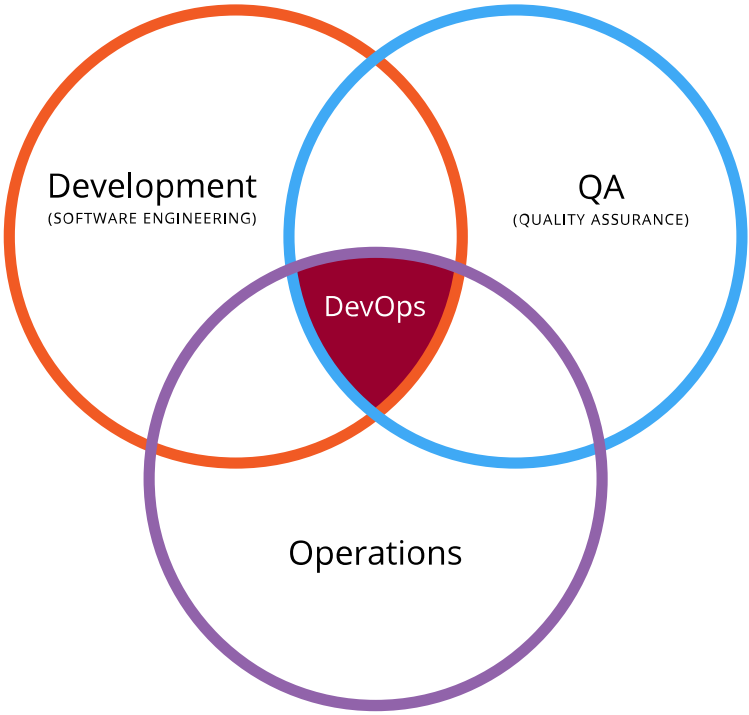
DevOps and DataOps

DevOps

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.

DevOps Essentials

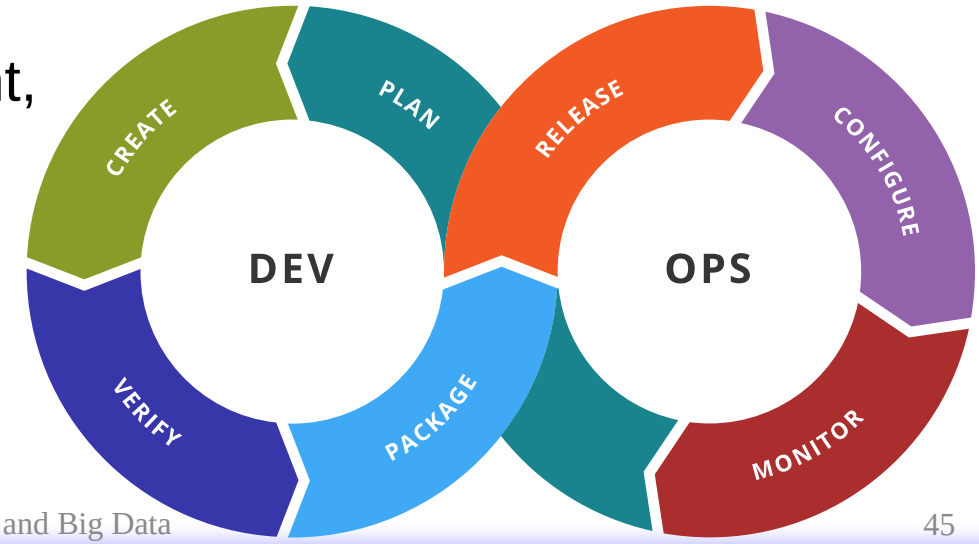
- Better Software, Faster time to market
- Movement Comes from Open Source
- Synergy of **Development and Operations**
- Covers the *entire* Application LifeCycle



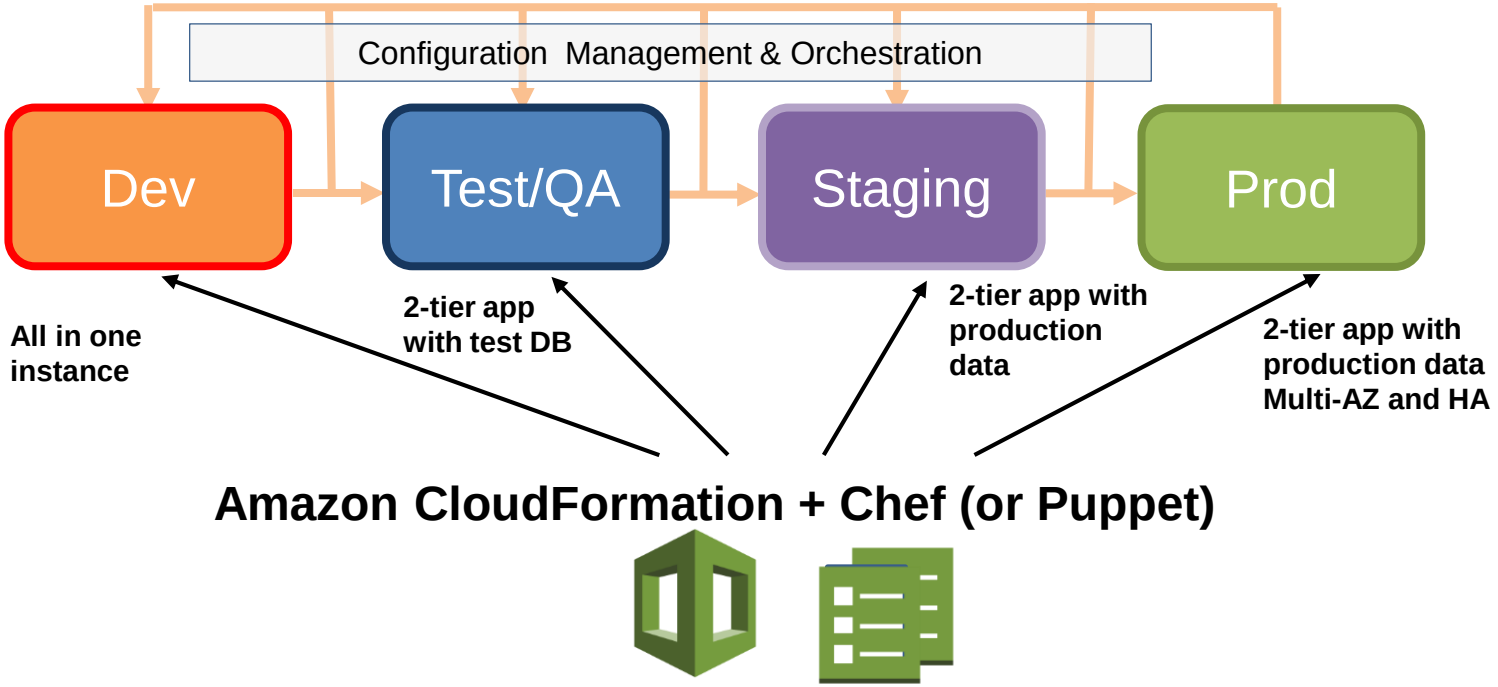
DevOps Toolchain

DevOps is a cultural shift and collaboration between development, operations and testing, enabled by DevOps toolchain

- **Code** — Code development and review, version control tools, code merging
- **Build** — Continuous integration tools, build status
- **Test** — Test and results determine performance
- **Package** — Artifact repository, application pre-deployment staging
- **Release** — Change management, release approvals, release automation
- **Configure** — Infrastructure configuration and management, Infrastructure as Code tools
- **Monitor** — Applications performance monitoring, end–user experience

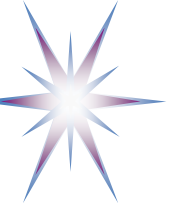


Cloud-powered Services Development Lifecycle



- Easily creates test environment close to real
- Powered by cloud deployment automation tools
  - To enable configuration Management and Orchestration, Deployment automation
- Continuous development – test – integration
  - CloudFormation Template, Configuration Template, Bootstrap Template
- Can be used with Puppet and Chef, two configuration and deployment management systems for clouds

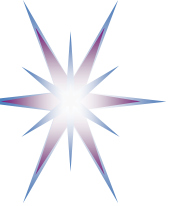
[ref] Building Powerful Web Applications in the AWS Cloud" by Louis Columbus
http://softwarestrategiesblog.com/2011/03/10/building-powerful-web-applications-in-the-aws-cloud/



Azure DevOps Services
(since 10 Sept 2018, former VSTS)

Azure DevOps Services is a cloud service for collaborating on code development.

- Azure Pipelines
  - CI/CD that works with any language, platform, and cloud.
- Azure Repos
  - Unlimited cloud-hosted private Git and TFVC repos for your project.
- Azure Boards
  - Work tracking with Kanban boards, backlogs, team dashboards, and custom reporting.
- Azure Test Plans
  - All-in-one planned and exploratory testing solution.
- Azure Artifacts
  - Maven, npm, and NuGet package feeds from public and private sources.
- Built-in wiki for sharing information with DevOps team



Azure DevOps Processes

Continuous integration (CI)

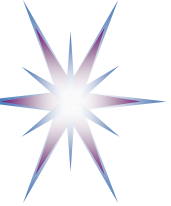
- Take advantage of continuous integration to improve software development quality and speed. When you use Azure DevOps or Jenkins to build apps in the cloud and deploy to Azure, each time you commit code, it's automatically built and tested—so bugs are detected faster.

Continuous delivery (CD)

- Ensure that code and infrastructure are always in a production-deployable state, with continuous delivery. By combining continuous integration and infrastructure as code (IaC), you'll achieve identical deployments and the confidence you need to manually deploy to production at any time.

Continuous deployment with CI/CD

- With continuous deployment, you can automate the entire process from code commit to production if your CI/CD tests are successful. Using CI/CD practices, paired with monitoring tools, you'll be able to safely deliver features to your customers as soon as they're ready.



DataOps

- DataOps (data operations) is an emerging discipline that brings together DevOps teams with data engineer and data scientist roles to provide the tools, processes and organizational structures to support the data-focused enterprise.
- DataOps is a new approach to the end-to-end data lifecycle, which applies new processes and methodologies to data analytics.
- Agile software development helps deliver new analytics faster and with higher quality.
- DevOps automates the deployment of new analytics and data.
- Statistical process controls, used in lean manufacturing, test and monitor the quality of data flowing through the data-analytics pipeline.

Data Exchange and Data Markets

Data Exchange and Data Markets

- IoT is considered as a key use case and a facilitator for Data Markets
  - Potentially many consumers for centrally or locally operated IoT infrastructure
  - IoT networks create valuable data that can be used for multiple purpose
  - IoT data can be exchanged and traded
- Open and Public data
- Current trading models are simple and in most cases are free or by subscription
- Making data economical goods would require new operational models, infrastructure and tools

Modern data architecture vs Data Market

Characteristics of modern data architecture

1. Customer-centric
2. Automated
3. Smart
4. Adaptable, Agile
5. Cloud based, Elastic
6. Collaborative
7. Governed
8. Secure, Trusted

Characteristics of emerging data markets

1. Customer-centric
2. Automated
3. Smart
4. Regional/sectoral specialised
5. Cloud powered/integrated
6. Collaborative
7. Governed
8. Secure, Trusted
9. Auditable
10. Transparent
11. Commoditised/Monetised
12. Combining data and algorithms (as part of containers)

Data Properties as Economic Goods

STREAM data principles for industrial and commoditised data

- [S] Sovereign
- [T] Trusted
- [R] Reusable
- [E] Exchangeable
- [A] Actionable
- [M] Measurable

- Other data properties: Important ***to commoditise*** data
  - Quality, Valuable, Auditable/Trackable, Brandable, Authentic
  - Interoperable, Findable, Accessible, not-Rival, Composable
  - Ownership and IPR

- Leverages FAIR principles for research data
  - Findable – Accessible – Interoperable - Reusable

Data Market Architecture components

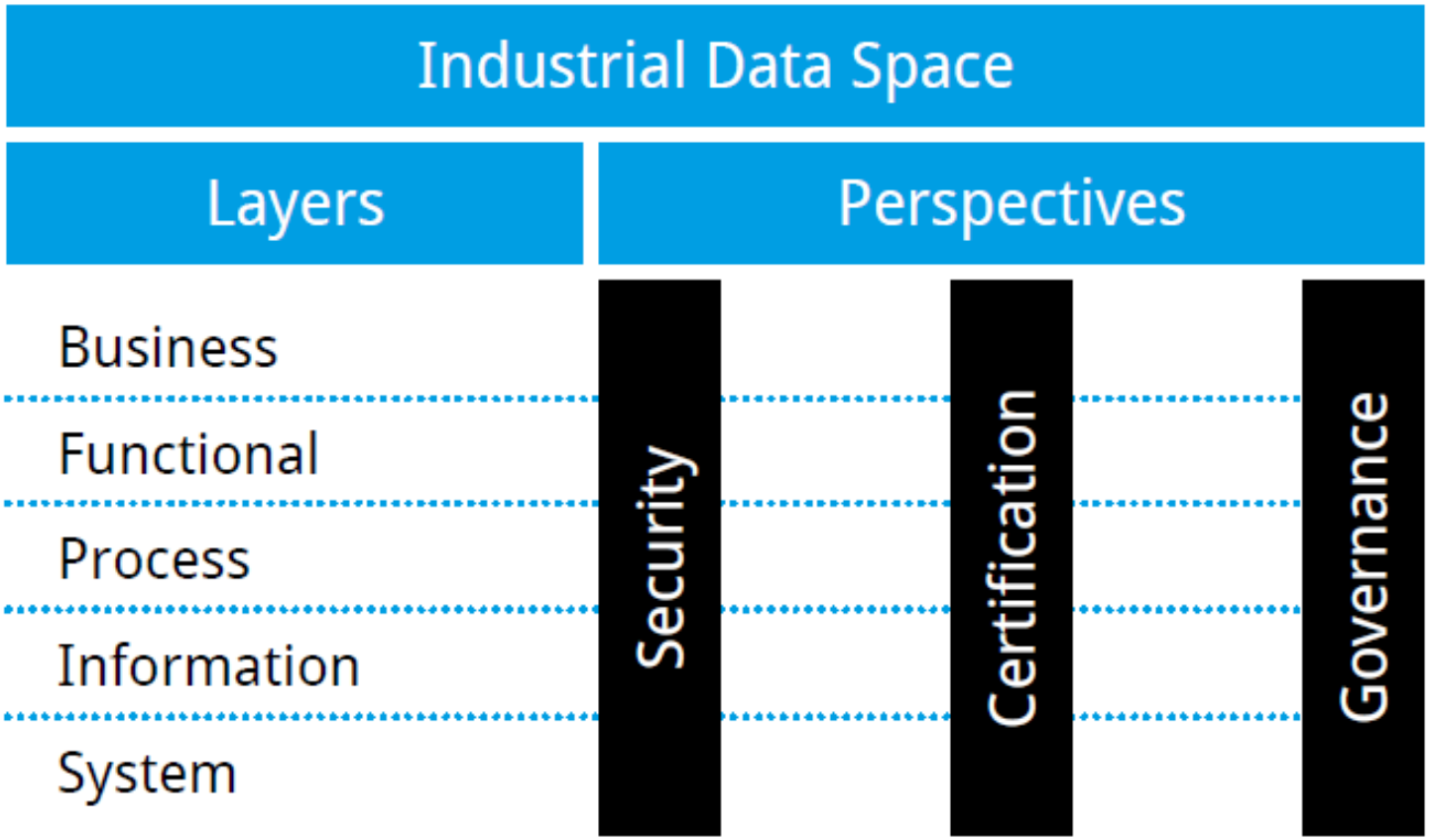
- **Data Source** (producer/seller/publisher) – Supply side
- **Data Target** (consumer, buyer, subscriber) – Demand side
- **Data Broker**
  - Data/Value Broker
  - Trust Broker or Trusted Introducer
- **Directory or Catalog service**
  - Including data (quality) ranking
  - Including API link or repository
- **Data Exchange**
  - Infrastructure component vs peer-to-peer customer network
  - Provenance and transactions control issue
- **Data Storage**/Cache Data Delivery Network
  - Data Lake (HDFS based and SQL/NoSQL)
  - Caching for traffic optimisation in DDN
- **Open/Public Data** access and storage
  - Can be stored to offer as part of DM storage
  - Facilitate quality of data analytics
- Data Transfer vs Data Access
  - Scenario: App container vs Data container
  - Container security using Intel TXT or SGX technology
- (Optional) Secure data processing
  - Used for data quality inspection (e.g. P.I.D., or IP), auditing, composition
  - Data preparation/conditioning

Re-using experience of Internet eXchange, Cloud eXchange and Financial Exchanges

International Data Space Association

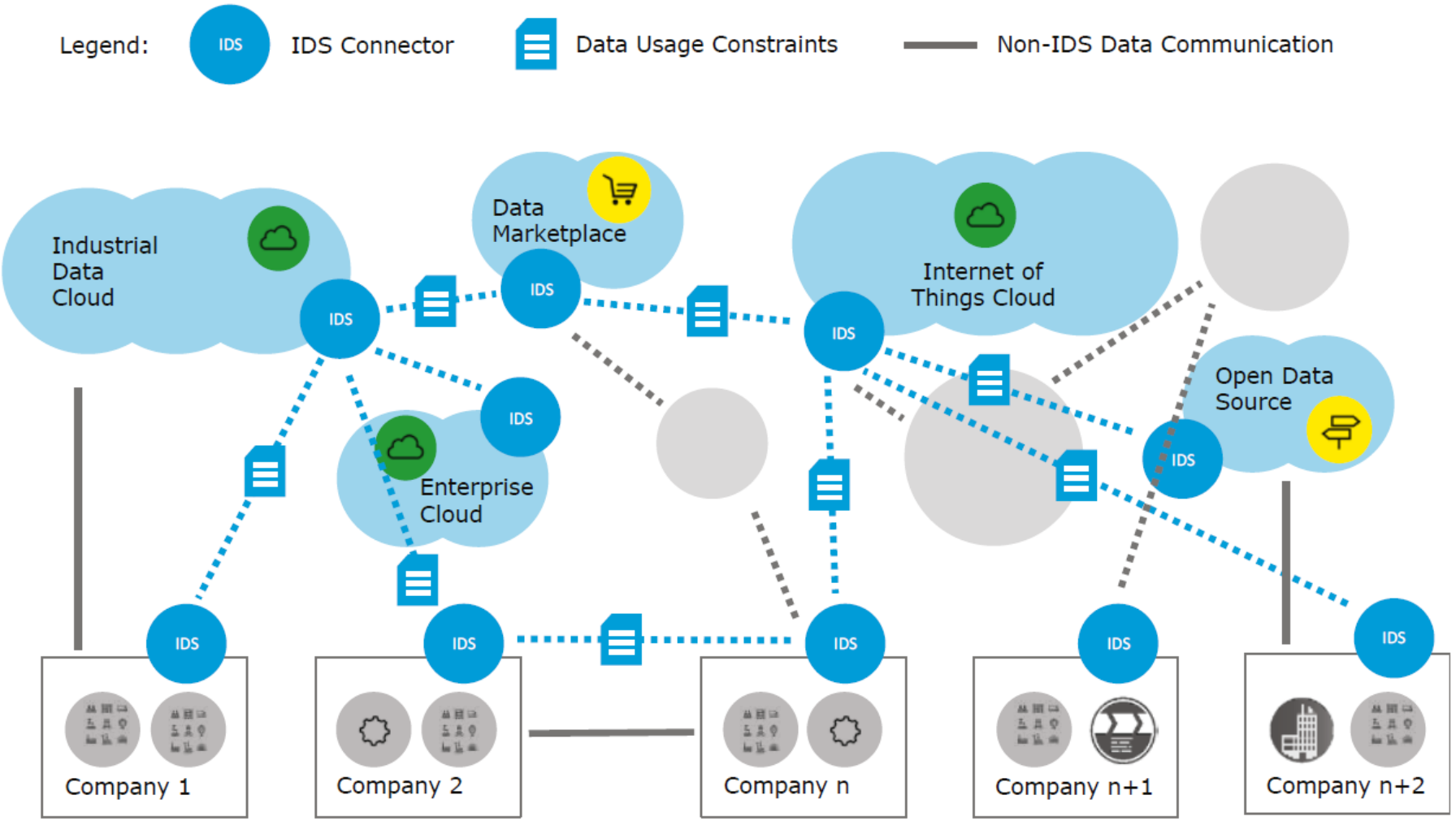
- Started 2016 as Industrial Data Space initiative (supported by German project)
- Re-defined as International Data Space Association (IDSA)
  - Published International Data Space Architecture Version 2.0 (2018)
  - Whitepaper and use cases
- Associated H2020 projects
  - Boost4.0 – Big Data for Factories (20 Mln (100 Mln private), 3yrs, 50 partners, 16 countries)
  - MIDIH – Manufacturing Industry Digital Innovation Hub (22 partners, 12 countries)
    - Services: technological, business, skills building
    - Open calls
  - Close cooperation with FIWARE Foundation (cloud like infrastructure resulted from Future Internet program)
    - Positions itself against IoT and Open-Data solutions in the areas of smart cities, Industry 4.0 and agriculture
- Ongoing active outreach developments
  - Data Sovereignty and Secure Data Exchange model
  - Data Markets
  - Trusted platforms

General Structure of IDS Architecture

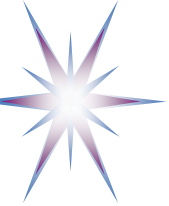


- Specification defines functionalities by layers
- Details are sufficient to define processes, functional components and API

Cloud based IDS infrastructure for Data Exchange and Trading



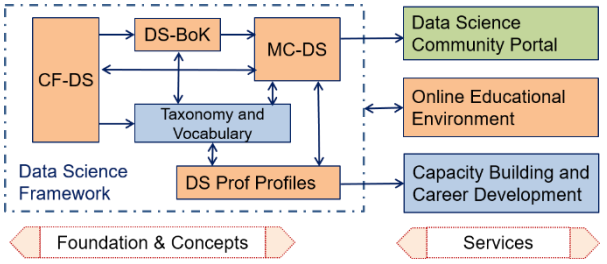
- IDS Connector is the main functional component
- No specifically defined common infrastructure services

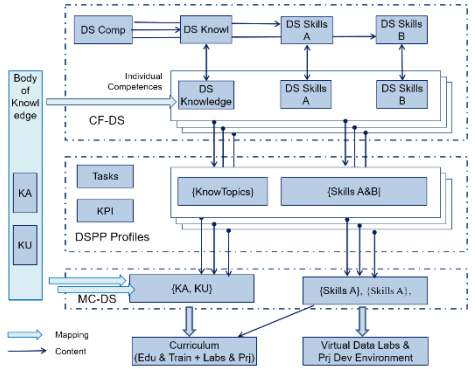


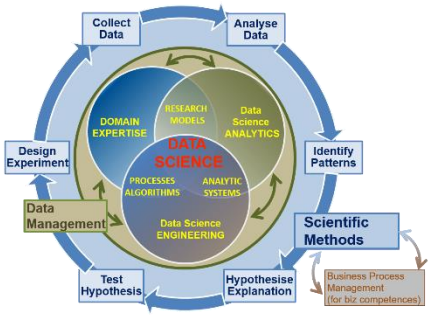
Competences, Skills and Learning

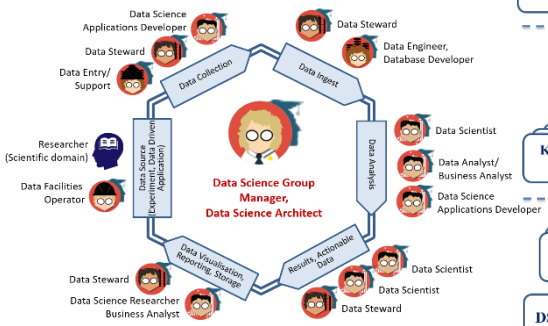
EDISON Products for Data Science Skills Management and Curriculum Design

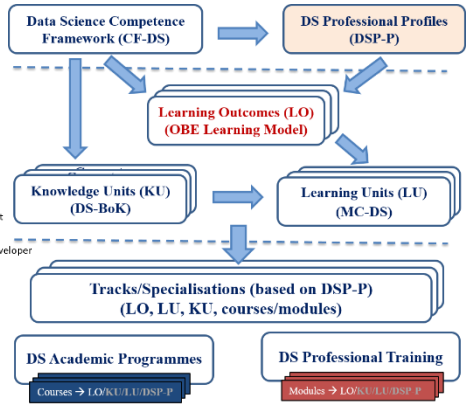
- EDISON Data Science Framework (EDSF)
  - Compliant with EU standards on competences and professional occupations e-CFv3.0, ESCO
  - Customisable courses design for targeted education and training
- Skills development and career management for Core Data Experts and related data handling professions
- Capacity building and Data Science team design
- Academic programmes and professional training courses (self) assessment and design
- Cooperation with International professional organisations IEEE, ACM, BHEF, APEC (AP Economic Cooperation)



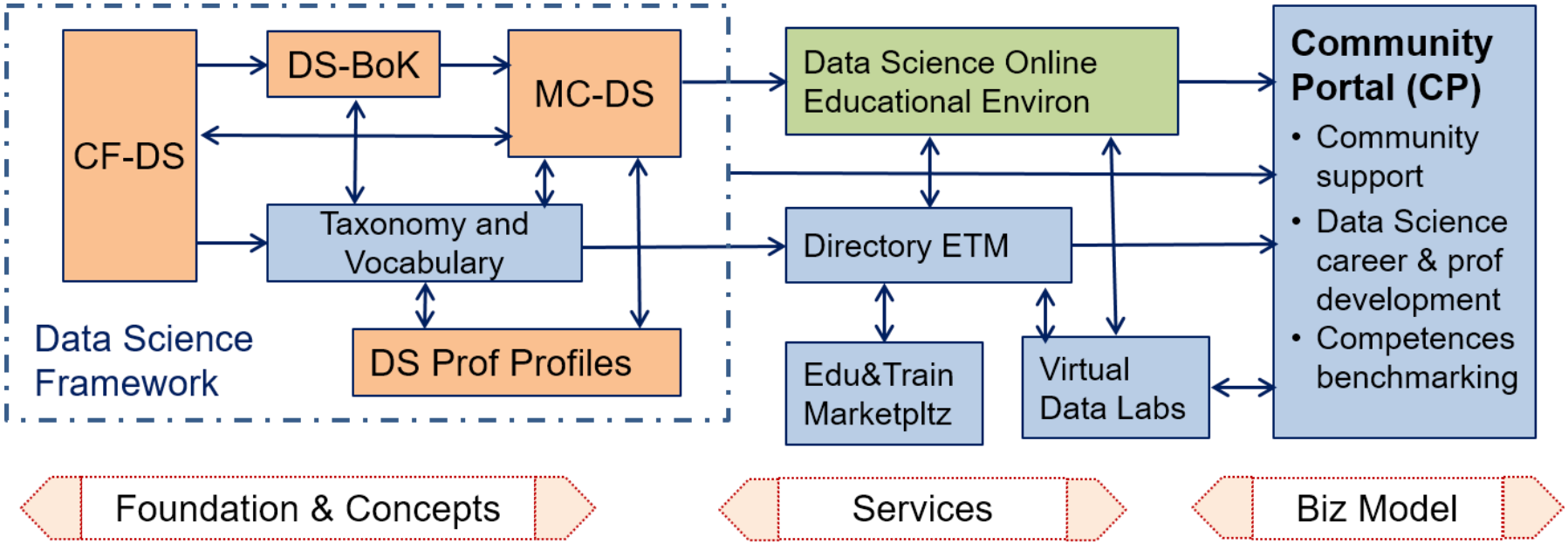








EDISON Data Science Framework (EDSF)



EDISON Framework components

- CF-DS – Data Science Competence Framework
- DS-BoK – Data Science Body of Knowledge
- MC-DS – Data Science Model Curriculum
- DSP – Data Science Professional profiles
- Data Science Taxonomies and Scientific Disciplines Classification
- EOEE - EDISON Online Education Environment

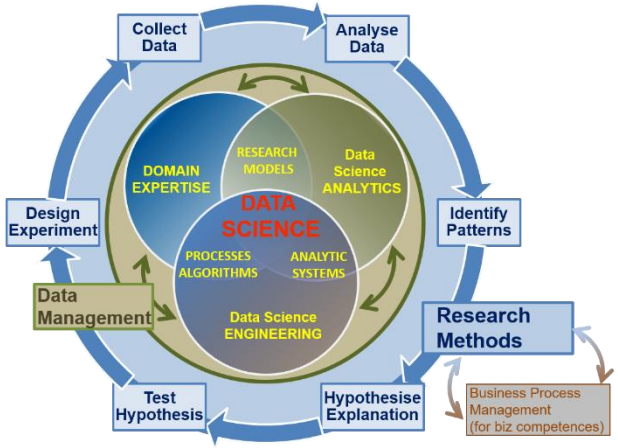
Methodology

- ESDF development based on job market study, existing practices in academic, research and industry.
- Review and feedback from the ELG, expert community, domain experts.
- Input from the champion universities and community of practice.

Data Scientist definition

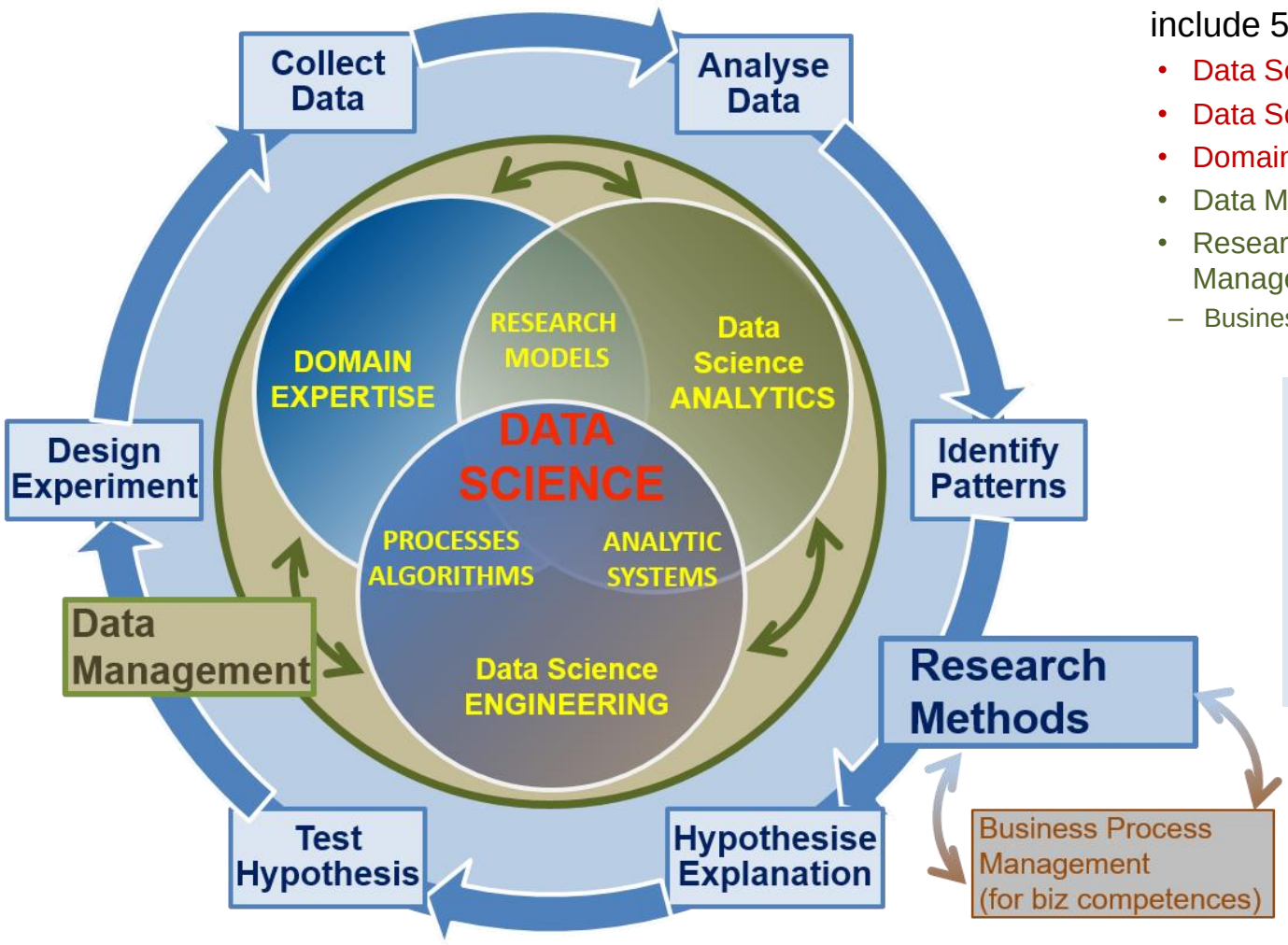
Based on the definitions by NIST SP1500 – 2015, extended by EDISON

- *A **Data Scientist** is a practitioner who has sufficient knowledge in the overlapping regimes of expertise in **business needs, domain knowledge, analytical skills, and programming and systems engineering expertise** to manage the end-to-end scientific method process through each stage in the **big data lifecycle** till the delivery of an **expected scientific and business value** to organisation or project.*



- Core Data Science competences and skills groups
  - **Data Science Analytics** (including Statistical Analysis, Machine Learning, Business Analytics)
  - **Data Science Engineering** (including Software and Applications Engineering, Data Warehousing, Big Data Infrastructure and Tools)
  - **Domain Knowledge and Expertise** (Subject/Scientific domain related)
- EDISON identified 2 additional competence groups demanded by organisations
  - **Data Management, Data Governance, Stewardship, Curation, Preservation**
  - **Research Methods and/vs Business Processes/Operations**
- **Data Science professional skills**: Thinking and acting like Data Scientist – required to successfully develop as a Data Scientist and work in Data Science teams

Data Science Competence Groups - Research



Data Science Competences include 5 groups

- Data Science Analytics
- Data Science Engineering
- Domain Knowledge and Expertise
- Data Management
- Research Methods and Project Management
  - Business Process Management (biz)

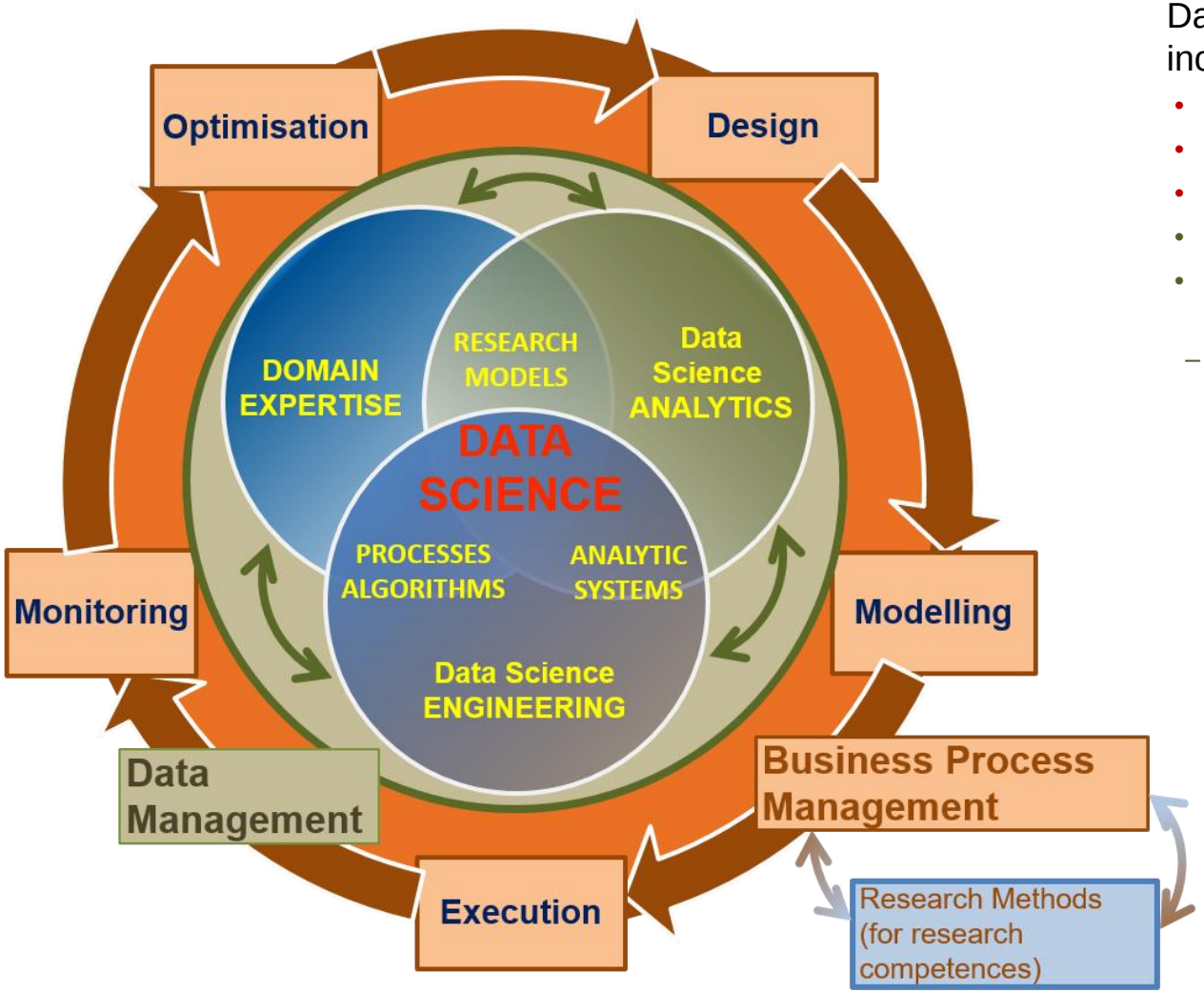
Scientific Methods

- Design Experiment
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesis Explanation
- Test Hypothesis

Business Operations

- Operations Strategy
- Plan
- Design & Deploy
- Monitor & Control
- Improve & Re-design

Data Science Competences Groups – Business



Data Science Competences include 5 groups

- Data Science Analytics
- Data Science Engineering
- Domain Knowledge and Expertise
- Data Management
- Research Methods and Project Management
  - Business Process Management (biz)

Scientific Methods

- Design Experiment
- Collect Data
- Analyse Data
- Identify Patterns
- Hypothesise Explanation
- Test Hypothesis

Business Process Operations/Stages

- Design
- Model/Plan
- Deploy & Execute
- Monitor & Control
- Optimise & Re-design

Identified Data Science Competence Groups

	Data Science Analytics (DSDA)	Data Science Engineering (DSENG)	Data Management and Governance (DSDM)	Research/Scientific Methods and Project Management (DSRMP)	Data Science Domain Knowledge, e.g. Business Analytics (DSDK/DSBPM)
0	Use appropriate data analytics and statistical techniques on available data to deliver insights into research problem or org. processes and support decision making	Use engineering principles and modern computer technology to research, design, implement new data analytics applications, develop experiments, processes, instruments, systems and infrastructures to support data handling during the whole data lifecycle	Develop and implement data management strategy for data collection, storage, preservation, and availability for further processing.	Create new understandings and capabilities by using the scientific method (hypothesis, test/artefact, evaluation) or similar engineering methods to discover new approaches to create new knowledge and achieve research or organisational goals	DSDK/DSBA Use domain knowledge (scientific or business) to develop relevant data analytics applications; adopt general Data Science methods to domain specific data types and presentations, data and process models, organisational roles and relations
1	**DSDA01** Effectively use variety of data analytics techniques	**DSENG01** Use engineering principles (general and software) to research, design, develop and implement new instruments and applications	DSDM01 Develop and implement data strategy, in particular, Data Management Plan (DMP)	**DSRMP01** Create new understandings and capabilities by using scientific/ research methods	**DSBPM01** Understand business and provide insight, translate unstructured business problems into an abstract mathematical framework
2	**DSDA02** Apply designated quantitative techniques	**DSENG02** Develop and apply computer methods to domain related problems	DSDM02 Develop data models including metadata	**DSRMP02** Direct systematic study toward a fuller knowledge or understanding of the observable facts	**DSBPM02** Participate strategically and tactically in financial decisions
3	**DSDA03** Pull together data from diff sources …	**DSENG03** Develop and prototype data analytics applications	DSDM03 Collect integrate data	**DSRMP03** Undertakes creative work	**DSBPM03** Provides support services to other
4	**DSDA04** Use diff perform techniques	**DSENG04** Develop, deploy operate Big Data storage	DSDM04 Maintain repository	**DSRMP04** Translate strategies into actions	**DSBPM04** Analyse data for marketing
5	**DSDA05** Develop analytics applic	**DSENG05** Apply security mechanisms	DSDM05 Visualise cmplx data	**DSRMP05** Contribute to organis goals	**DSBPM05** Analyse optimise customer relatio
6	**DSDA06** Visualise results of analysis, dashboards	**DSENG06** Design, build, operate SQL and NoSQL	**DSRM06** Develop and manage prolicies	**DSRMP06** Develop and guide data driven projects	**DSBPM06** Analyse data for marketing

Identified Data Science *Skills/Experience* Groups

Skills Type A – Based on knowledge acquired

- **Group 1: Skills/experience related to competences**
  - Data Analytics and Machine Learning
  - Data Management/Curation (including both general data management and scientific data management)
  - Data Science Engineering (hardware and software) skills
  - Scientific/Research Methods or Business Process Management
  - Application/subject domain related (research or business)
- **Group 2: Mathematics and statistics**
  - Mathematics and Statistics and others

Skills Type B – Base on practical or workplace experience

- **Group 3: Big Data (Data Science) tools and platforms**
  - Big Data Analytics platforms
  - Mathematics & Statistics applications & tools
  - Databases (SQL and NoSQL)
  - Data Management and Curation platform
  - Data and applications visualisation
  - ***Cloud based platforms and tools***
- **Group 4: Data analytics programming languages and IDE**
  - General and specialized development platforms for data analysis and statistics
- **Group 5: Soft skills and Workplace skills**
  - Data Science professional skills: Thinking and Acting like Data Scientist
  - 21st Century Skills: Personal, inter-personal communication, team work, professional network

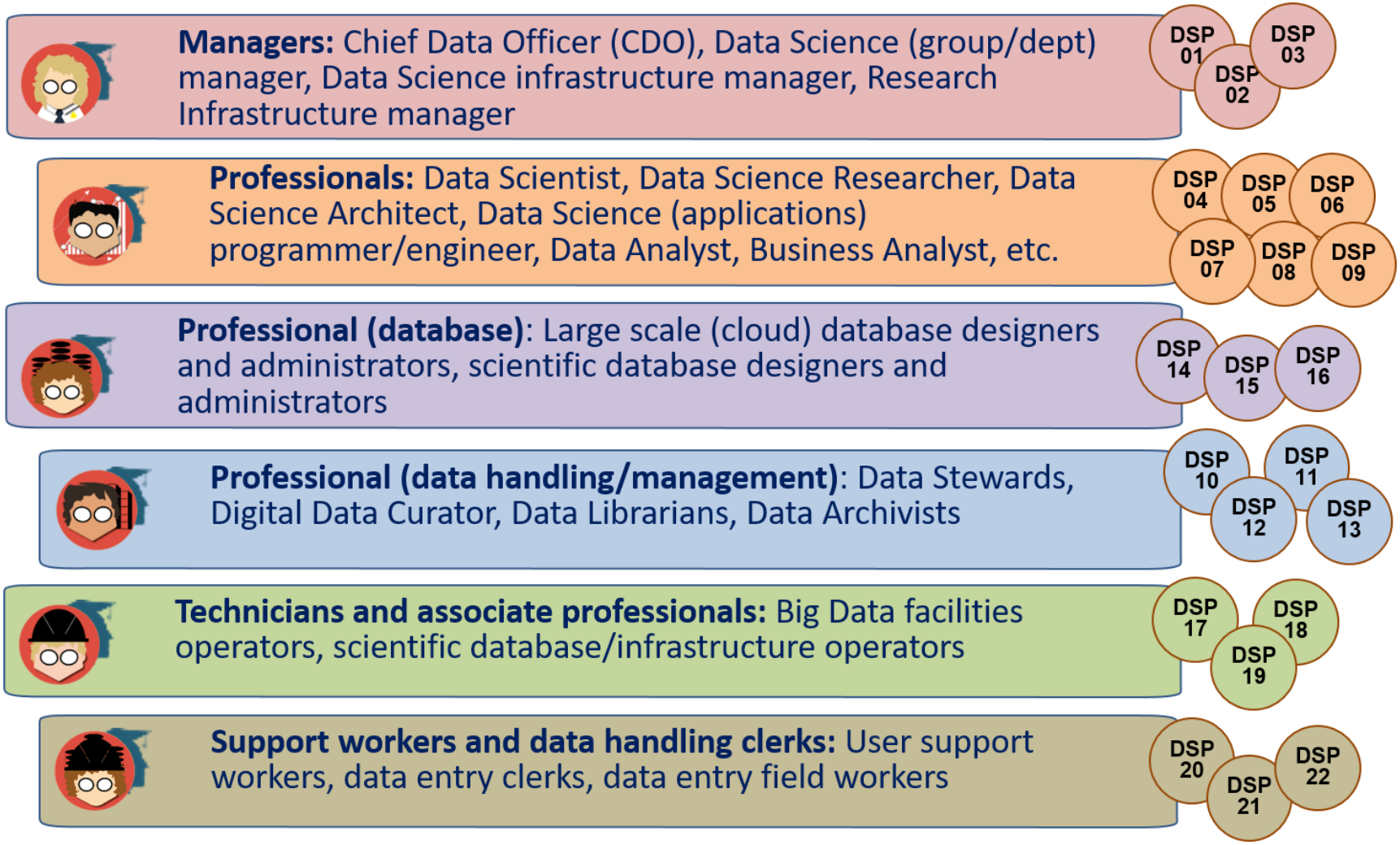
Data Science Professional Skills: Thinking and Acting like Data Scientist

1. **Recognise value of data**, work with raw data, exercise good data intuition, use SN and open data
2. Accept (be ready for) **iterative development**, know when to stop, comfortable with failure, accept the symmetry of outcome (both positive and negative results are valuable)
3. Good **sense of metrics**, understand importance of the results validation, never stop looking at individual examples
4. **Ask the right questions**
5. **Respect domain/subject matter knowledge** in the area of data science
6. **Data driven problem solver** and **impact-driven mindset**
7. **Be aware about power and limitations** of the main machine learning and data analytics algorithms and tools
8. Understand that most of **data analytics algorithms are statistics and probability based**, so any answer or solution has some degree of probability and represent an optimal solution for a number variables and factors
9. Recognise what things are **important** and what things are **not important** (in data modeling)
10. Working in **agile environment** and coordinate with other roles and team members
11. Work in **multi-disciplinary team**, ability to communicate with the domain and subject matter experts
12. Embrace **online learning**, continuously improve your knowledge, use **professional netw**orks and communities
13. **Story Telling**: Deliver actionable result of your analysis
14. **Attitude**: Creativity, curiosity (willingness to challenge status quo), commitment in finding new knowledge and progress to completion
15. **Ethics and responsible use** of data and insight delivered, awareness of dependability (data scientist is a feedback loop in data driven companies)

21st Century Skills (DARE & BHEF & EDISON)

1. **Critical Thinking:** Demonstrating the ability to apply critical thinking skills to solve problems and make effective decisions
2. **Communication:** Understanding and communicating ideas
3. **Collaboration:** Working with other, appreciation of multicultural difference
4. **Creativity and Attitude:** Deliver high quality work and focus on final result, intitiative, intellectual risk
5. **Planning & Organizing:** Planning and prioritizing work to manage time effectively and accomplish assigned tasks
6. **Business Fundamentals:** Having fundamental knowledge of the organization and the industry
7. **Customer Focus:** Actively look for ways to identify market demands and meet customer or client needs
8. **Working with Tools & Technology:** Selecting, using, and maintaining tools and technology to facilitate work activity
9. **Dynamic (self-) re-skilling:** Continuously monitor individual knowledge and skills as shared responsibility between employer and employee, ability to adopt to changes
10. **Professional networking:** Involvement and contribution to professional network activities
11. **Ethics:** Adhere to high ethical and professional norms, responsible use of power data driven technologies, avoid and disregard un-ethical use of technologies and biased data collection and presentation

Data Science Professions Family

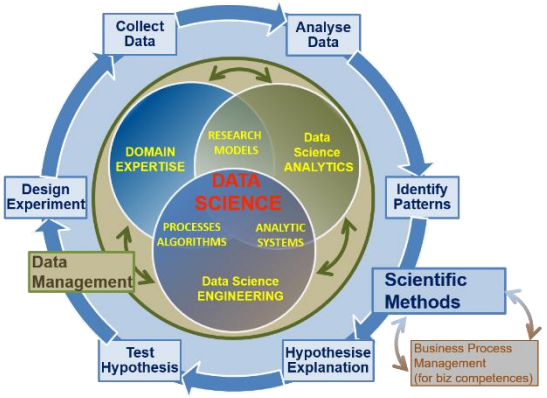


Icons used: Credit to [ref] https://www.datacamp.com/community/tutorials/data-science-industry-infographic

1. Data Science Body of Knowledge (DS-BoK)

What Data Scientists (different profiles) need to know

DS-BoK Knowledge Area Groups (KAG)



- KAG1-DSA: Data Analytics group including Machine Learning, statistical methods, and Business Analytics
- KAG2-DSE: Data Science Engineering group including Software and infrastructure engineering
- KAG3-DSDM: *Data Management group including data curation, preservation and data infrastructure*
- KAG4-DSRM: *Research Methods and Project Management group*
- KAG5-DSBA: Business Analytics and Business Intelligence

- KAG* - DSDK: Data Science domain knowledge to be defined by related expert groups

Data Science Body of Knowledge (1)

KA Groups	Suggested DS Knowledge Areas (KA)	Knowledge Areas from existing BoK and CCS2012 scientific subject groups
KAG1-DSDA: Data Science Analytics	KA01.01 (DSDA.01/SMDA) Statistical methods for data analysis KA01.02 (DSDA.02/ML) Machine Learning KA01.03 (DSDA.03/DM) Data Mining KA01.04 (DSDA.04/TDM) Text Data Mining KA01.05 (DSDA.05/PA) Predictive Analytics KA01.06 (DSDA.06/MODSIM) Computational modelling, simulation and optimisation	There is no formal BoK defined for Data Analytics. Data Science Analytics related scientific subjects from CCS2012: CCS2012: Computing methodologies CCS2012: Mathematics of computing CCS2012: Computing methodologies
KAG2-DSENG: Data Science Engineering	KA02.01 (DSENG.01/BDI) Big Data Infrastructure and Technologies KA02.02 (DSENG.02/DSIAPP) Infrastructure and platforms for Data Science applications KA02.03 (DSENG.03/CCT) Cloud Computing technologies for Big Data and Data Analytics KA02.04 (DSENG.04/SEC) Data and Applications security KA02.05 (DSENG.05/BDSE) Big Data systems organisation and engineering KA02.06 (DSENG.06/DSAPPD) Data Science (Big Data) applications design KA02.07 (DSENG.07/IS) Information systems (to support data driven decision making)	ACM CS-BoK selected KAs: AR - Architecture and Organization (including computer architectures and network architectures) CN - Computational Science IM - Information Management SE - Software Engineering (can be extended with specific SWEBOK KAs) SWEBOK selected KAs • Software requirements • Software design • Software engineering process • Software engineering models and methods • Software quality Data Science Analytics related scientific subjects from CCS2012

Data Science Body of Knowledge (2)

KA Groups	Suggested DS Knowledge Areas (KA)	Knowledge Areas from existing BoK and CCS2012 scientific subject groups
KAG3-DSDM: Data Management	KA03.01 (DSDM.01/DMORG) General principles and concepts in Data Management and organisation KA03.02 (DSDM.02/DMS) Data management systems KA03.03 (DSDM.03/EDMI) Data Management and Enterprise data infrastructure KA03.04 (DSDM.04/DGOV) Data Governance KA03.05 (DSDM.05/BDST0R) Big Data storage (large scale) KA03.06 (DSDM.05/DLIB) Digital libraries and archives	DM-BoK selected KAs (1) Data Governance, (2) Data Architecture, (3) Data Modelling and Design, (4) Data Storage and Operations, (5) Data Security, (6) Data Integration and Interoperability, (7) Documents and Content, (8) Reference and Master Data, (9) Data Warehousing and Business Intelligence, (10) Metadata, and (11) Data Quality.
KAG4-DSRM: Research Methods and Project Management	KA04.01 (DSRMP.01/RM) Research Methods KA04.01 (DSRMP.02/PM) Project Management	There are no formally defined BoK for research methods PMI-BoK selected KAs • Project Integration Management • Project Scope Management • Project Quality • Project Risk Management
KAG5-DSBPM: Business Analytics	KA05.01 (DSBA.01/BAF) Business Analytics Foundation KA05.02 (DSBA.02/BAEM) Business Analytics organisation and enterprise management	BABOK selected KAs *) Business Analysis Planning and Monitoring Requirements Life Cycle Management Solution Evaluation and improvements recommendation

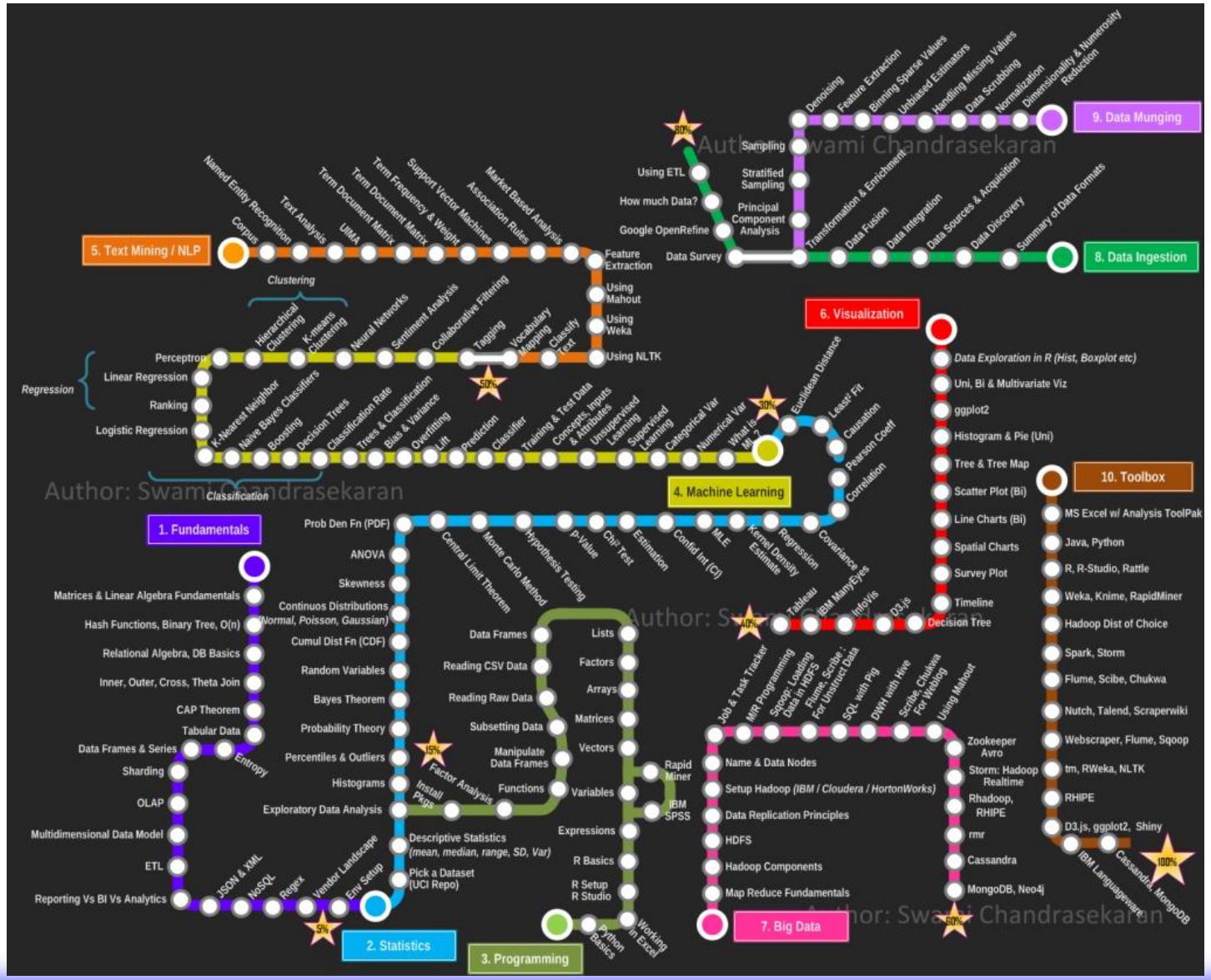
DSP04 Data Scientist – Required practical skills and Hands-on labs

Data Science curriculum should include the following elements to achieve necessary skills Type B:

- Python (or R) and corresponding data analytics libraries
- NoSQL and SQL Databases (Hbase, MongoDB, Cassandra, Redis, Accumulo, MS SQL, My SQL, PostgreSQL, etc.)
- Big Data Analytics platforms (Hadoop, Spark, Data Lakes, others)
- Real time and streaming analytics systems (Flume, Kafka, Storm)
- Kaggle competition, resources and community platform, including rich data sets, forum and computing resources
- Visualisation software (D3.js, Processing, Tableau, Julia, Raphael, etc.)
- Web API management and web scrapping
- Git versioning system as a general platform for software development
- Development Frameworks: Python, Java or C/C++, AJAX (Asynchronous Javascript and XML), D3.js (Data-Driven Documents), jQuery, others
- Cloud based Big Data and data analytics platforms and services, including large scale storage systems
  - Essential for workplace adjustment

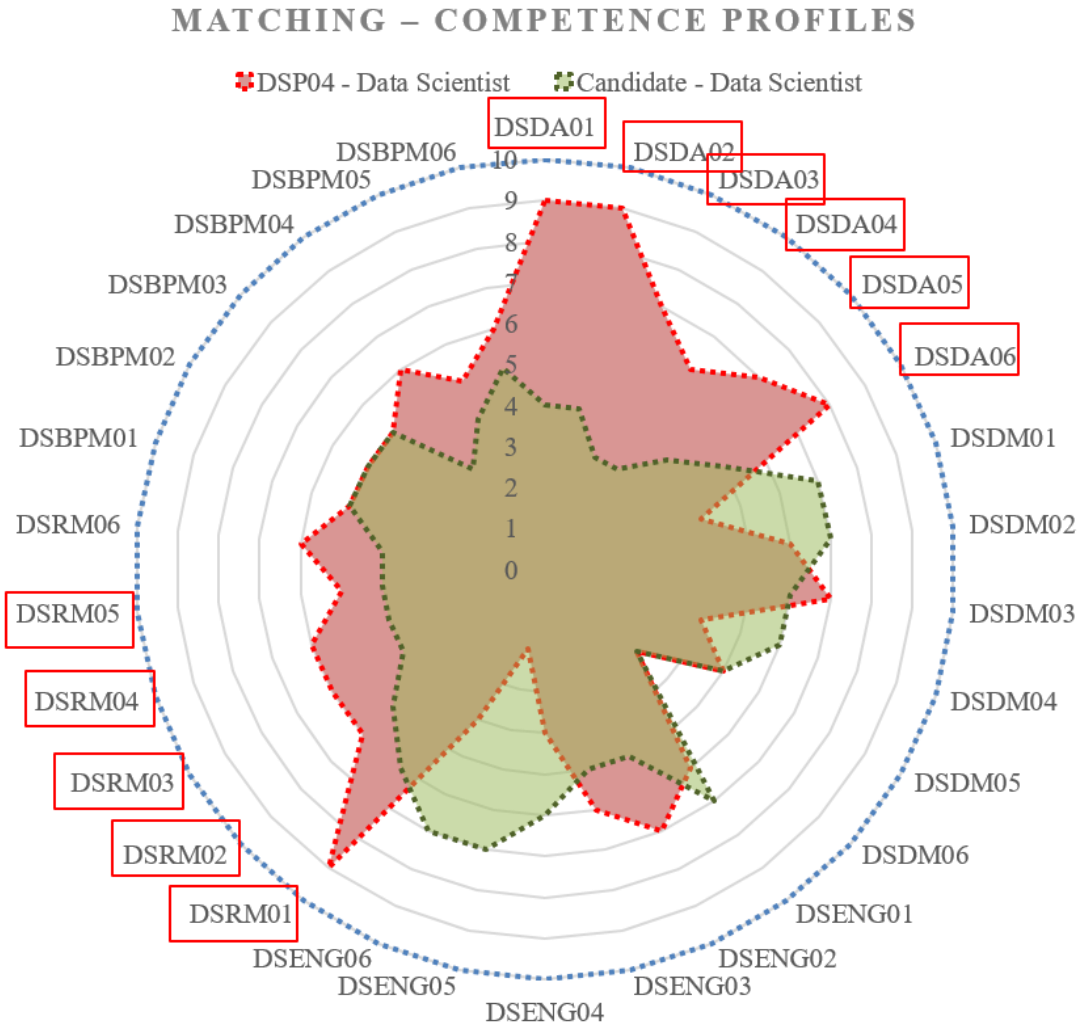
Becoming a Data Scientist by Swami Chandrasekaran (2013)

http://nirvacana.com/thoughts/becoming-a-data-scientist/



- Good and practical advice how to learn Data Science, step by step
- Follow the route

Individual Competences Benchmarking

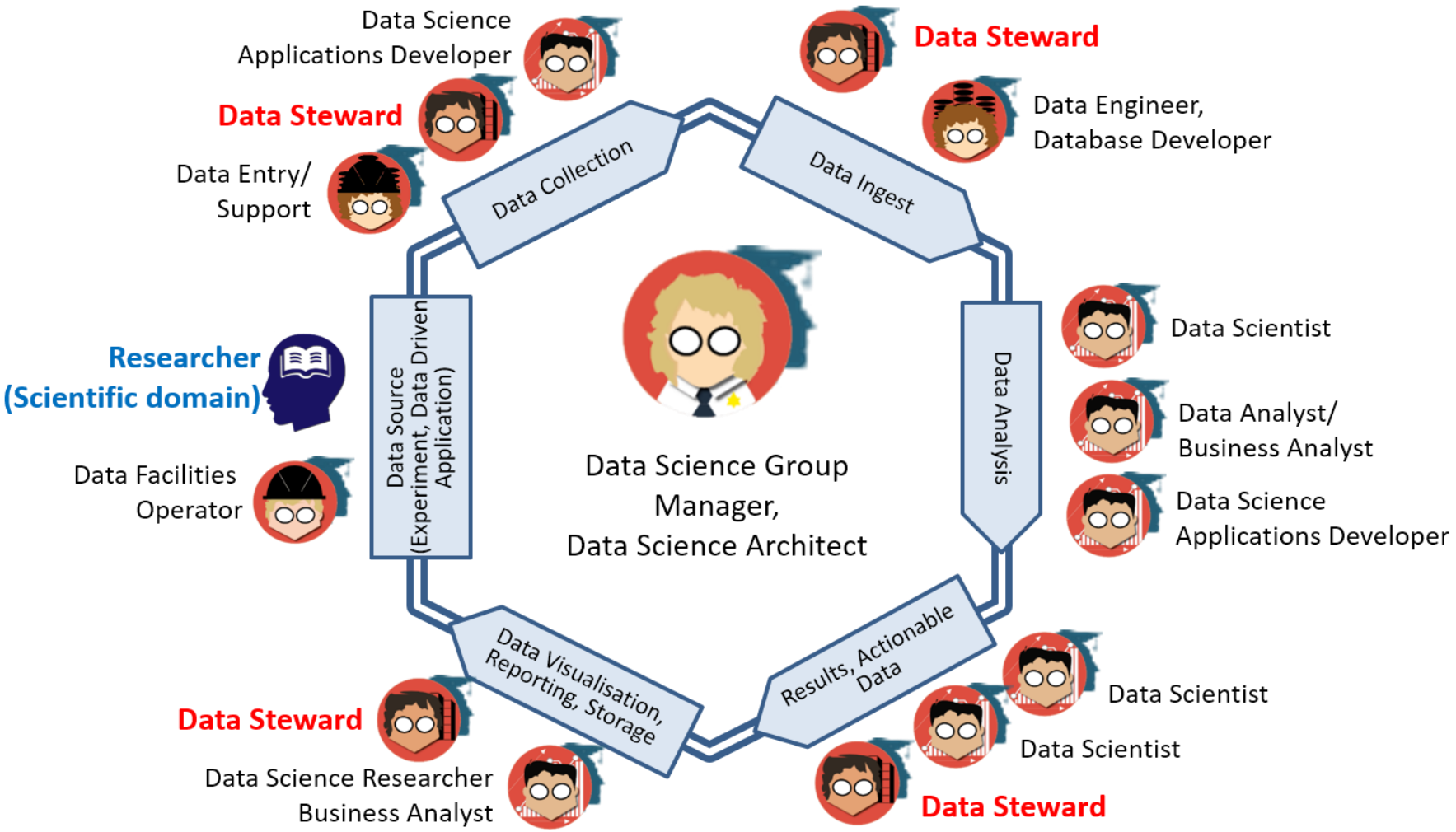


Individual Education/Training Path based on Competence benchmarking

- Red polygon indicates the chosen professional profile: Data Scientist (general)
- Green polygon indicates the candidate or practitioner competences/skills profile
- Insufficient competences (gaps) are highlighted in *red*
  - *DSDA01 – DSDA06 Data Science Analytics*
  - *DSRM01 – DSRM05 Data Science Research Methods*
- Can be use for team skills match marking and organisational skills management

[ref] For DSP Profiles definition and for enumerated competences refer to EDSF documents CF-DS and DSP Profiles.

Building a Data Science Team



Data Science or Data Management Group/Department: Organisational structure and staffing - EXAMPLE

Data Science or Data Management Group/Department

>> Reporting to CDO/CTO/CEO
- Providing cross-organizational services

- (Managing) Data Science Architect (1)
- Data Scientist (1), Data Analyst (1)
- Data Science Application programmer (2)
- Data Infrastructure/facilities administrator/operator: storage, cloud, computing (1)
- **Data stewards**, curators, archivists (3-5)

Estimated: Group of 10-12 data specialists for research institution of 200-300 research staff.

Growing role and demand for Data Stewards and data stewardship

Discussion: How to become a Data Scientist (or Big Data Architect)

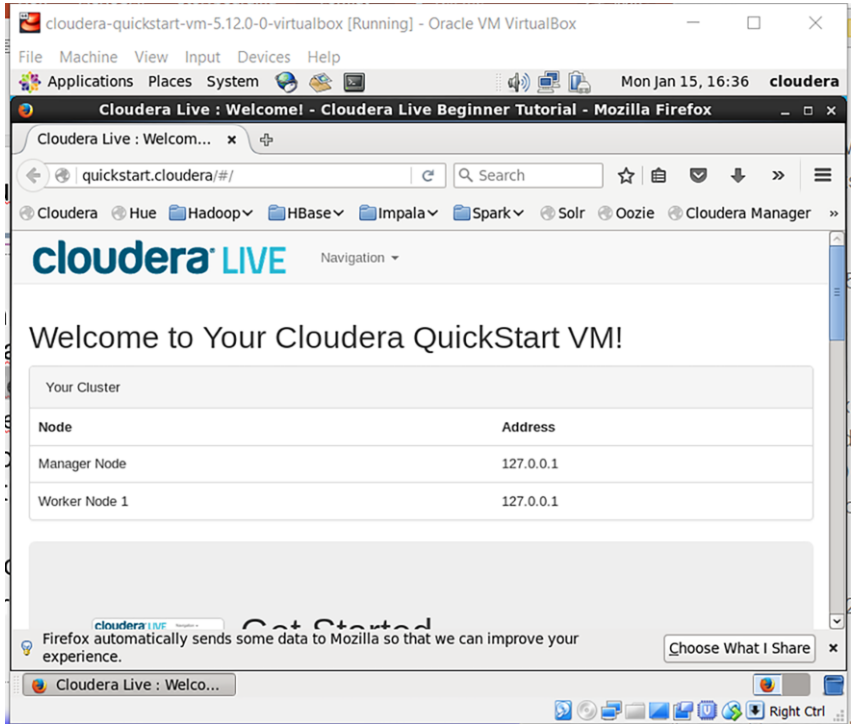
- Understand required Data Science and Analytics competences and skills
- Build your own learning path
  - Assess your knowledge and start from basics
  - Statistics is foundation of Data (Science) Analytics
    - Develop statistical/probabilistic thinking
    - Difference between Data Science and statistics
  - Learn from others experience: read blogs, join forums and communities
  - Decide about academic degree, professional certificate, self-education/training, join local Meetup
- Become involved in a real project
  - Remember variety of Data Scientist roles and profiles

Learning resources

- Microsoft Virtual Academy – FREE
- Google Qwik Lab - FREE
- Linkedin Learning – 1 month free -> 279 USD/Year
- Datacamp: Data Science with R and Python – 1 month free -> 380 USD/Year
- Coursera, Udacity
- Certification and training PMI, DAMA, IIBA

- Pre-configured Hadoop/Spark cluster for personal use
  - Cloudera QuickStart for VirtualBox and Docker
  - Hortonworks Sanbox on VM and on container
- Databricks Hadoop and Spark
  - Free community edition and services online
  - Databricks cluster available on Azure

Cloudera Quickstart VM for VirtualBox



Accounts

- Once you launch the VM, you are automatically logged in as the cloudera user. The account details are:
  - username: cloudera
  - password: cloudera
- The Cloudera account has sudo privileges in the VM. The root account password is cloudera.
- The root MySQL password (and the password for other MySQL user accounts) is also cloudera.
- Hue and Cloudera Manager use the same credentials.

Datasets

- Kaggle - https://www.kaggle.com/
- StackExchange data dump files https://data.stackexchange.com/stackoverflow/
- Google BigQuery and public datasets - https://cloud.google.com/public-datasets/
- AWS public datasets - https://aws.amazon.com/public-datasets/
- Azure public datasets Azure Analytics - https://docs.microsoft.com/en-us/azure/sql-database/sql-database-public-data-sets
- Azure DataMarket (retiring) http://datamarket.azure.com/browse
- IBM Watson Analytics Datasets - https://www.ibm.com/communities/analytics/watson-analytics-blog/guide-to-sample-datasets/
- The KB Lab hosts experimental tools and data sets based on the KB's digitised collection http://lab.kb.nl/
- Many others – check Wikipedia https://en.wikipedia.org/wiki/List_of_datasets_for_machine_learning_research

Grants possibilities for research and education

- AWS Education credits: for teachers and students (including w/o credit cards)
- AWS research grants – only for pre-production services
- Azure Research grants: 2-3 times a year on specified topics: ML, genomics, etc
- All major CSP has trial credits
  - AWS – 1 Year + 200 USD
  - Azure – 1 month + 200 USD
  - Google – 1 year + 300 USD

Development tools – Old and new

- Eclipse: Full cloud API integration
- CLI cloud API integrated with the OS platform CLI
- Azure Power Shell and Storage Explorer
- Variety of Cloud IDE/API pluggable modules for last 3 IDE
- Python tools and IDE: Jupyter/IPython Notebook, Anaconda, Spyder, Rodeo, Atom
- Google Compute Engine IDE and GO language IDE
- Visual Studio Code
- Visual Studio Community Edition
- Azure DevOps (limited free services)
- Atom Hackable Editor

Discussion and Questions

Suggested topics for discussion

- How to handle Big Data and Cloud complexity
- How to become a Data Scientist, Big Data Architect, etc