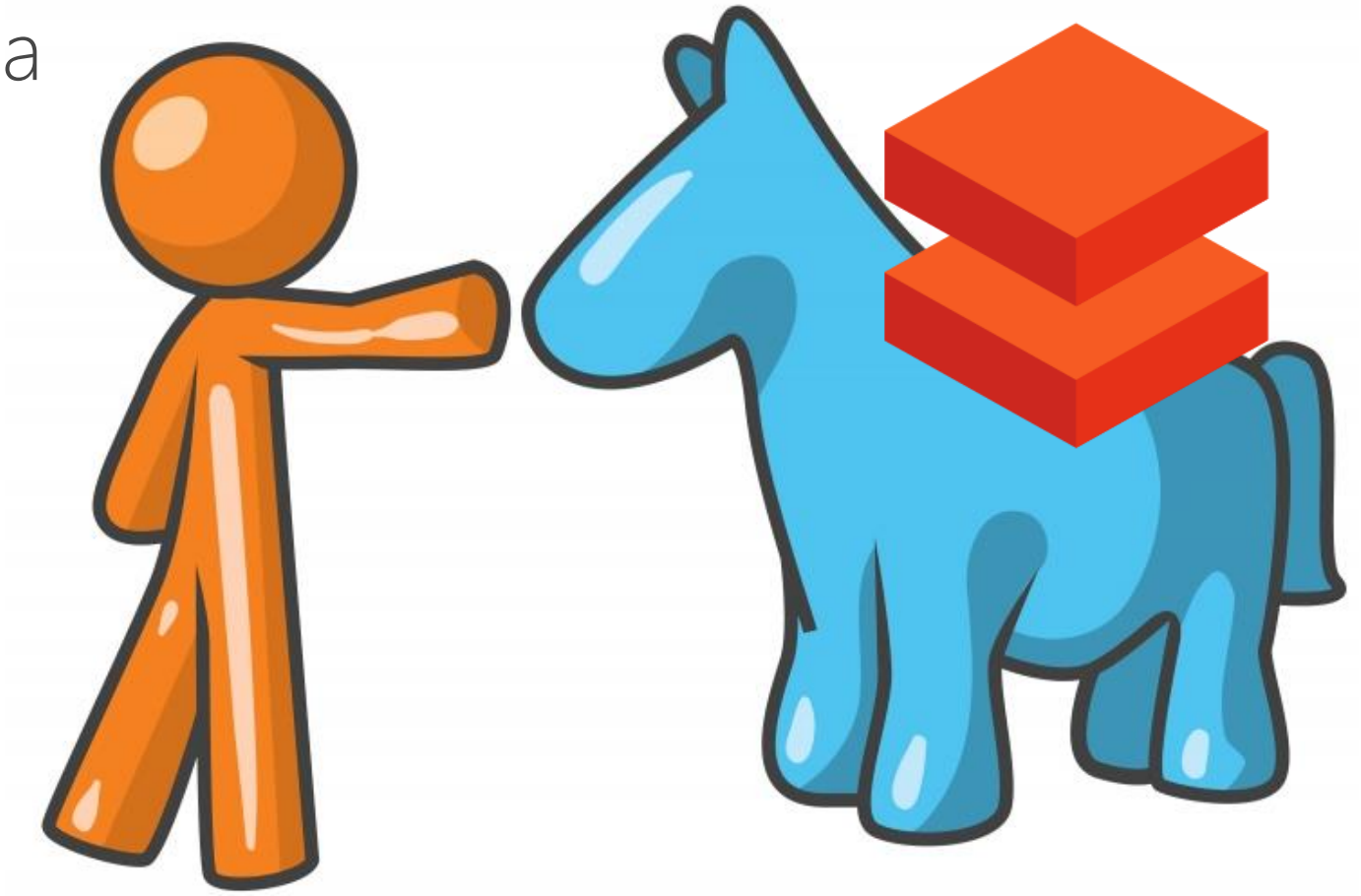
Azure Databricks

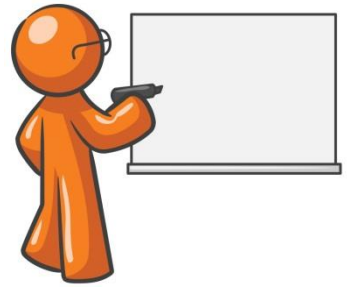
Den nye dreng i din Data Warehouse arkitektur

Hest?



ORANGEMAN

Agenda



- Data and AI People
- Introduction to Databricks in Azure
- Architecture
- Data processing with Azure Databricks
- Machine Learning with Azure Databricks
- Scale, Cost and Management
- Language, Data and Format
- Azure Databricks Delta

ORANGEMAN

DATA Engineer

DataCamp
Learn Data Science By Doing

DATA Scientist

Develops, constructs, tests, and maintains architectures. Such as databases and large-scale processing systems.

Cleans, massages and organizes (big) data. Performs descriptive statistics and analysis to develop insights, build models and solve a business need.

Data Scientist

also known as Data Managers, statisticians.

A data scientist will be able to take data science projects from end to end. They can help store large amounts of data, create predictive modelling processes and present the findings.

Skills: Mathematics, Programming, Communication

Will use programmes such as:

SQL, Python, R

Data Engineers

also known as database administrators and data architects.

They are versatile generalists who use computer science to help process large datasets. They typically focus on coding, cleaning up data sets, and implementing requests that come from data scientists.

Skills: Programming, Mathematics, Big data

Will use programmes such as:

Hadoop, NoSQL, and Python

Data Analysts

also known as business Analysts.

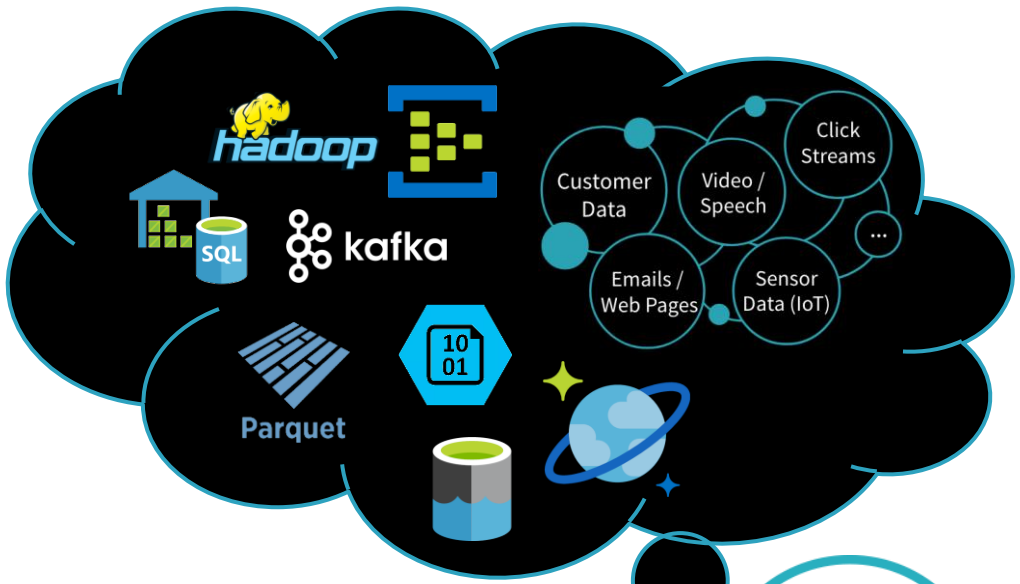
They typically help people from across the company understand specific queries with charts.

Skills: Statistics, Communication, Business knowledge

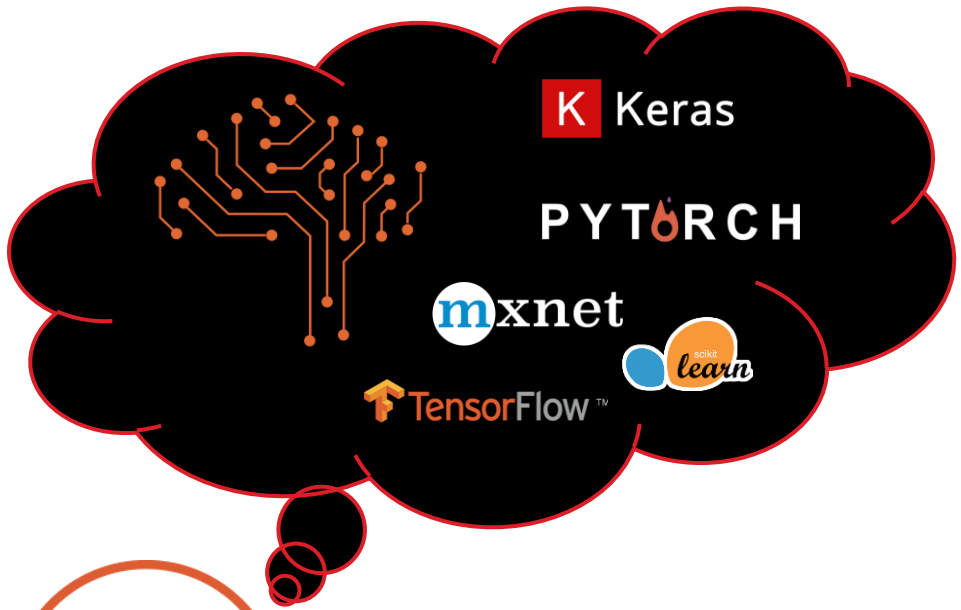
Will use programmes such as:

Excel, Tableau, SQL

Data & AI People are in Silos



DATA ENGINEERS

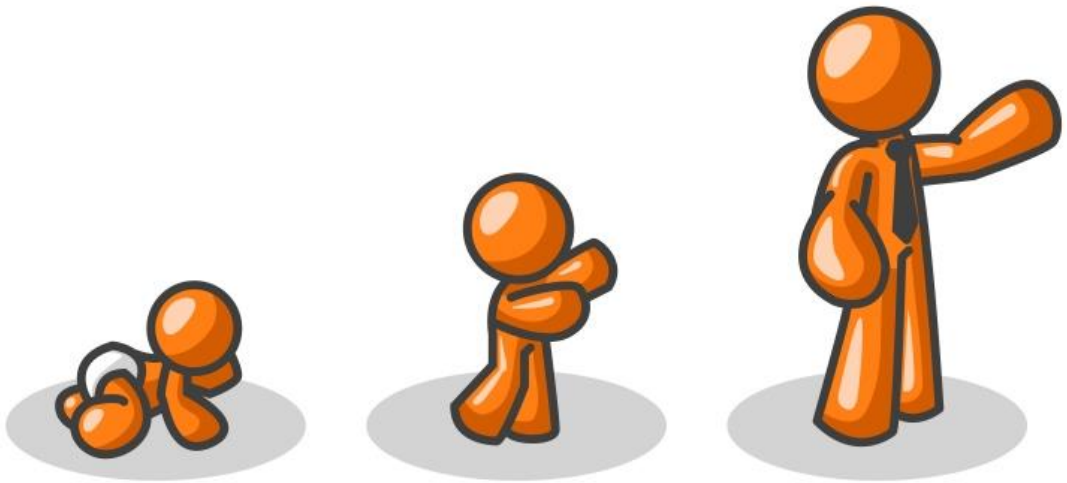


DATA SCIENTISTS

ORANGEMAN

Introduction to Databricks

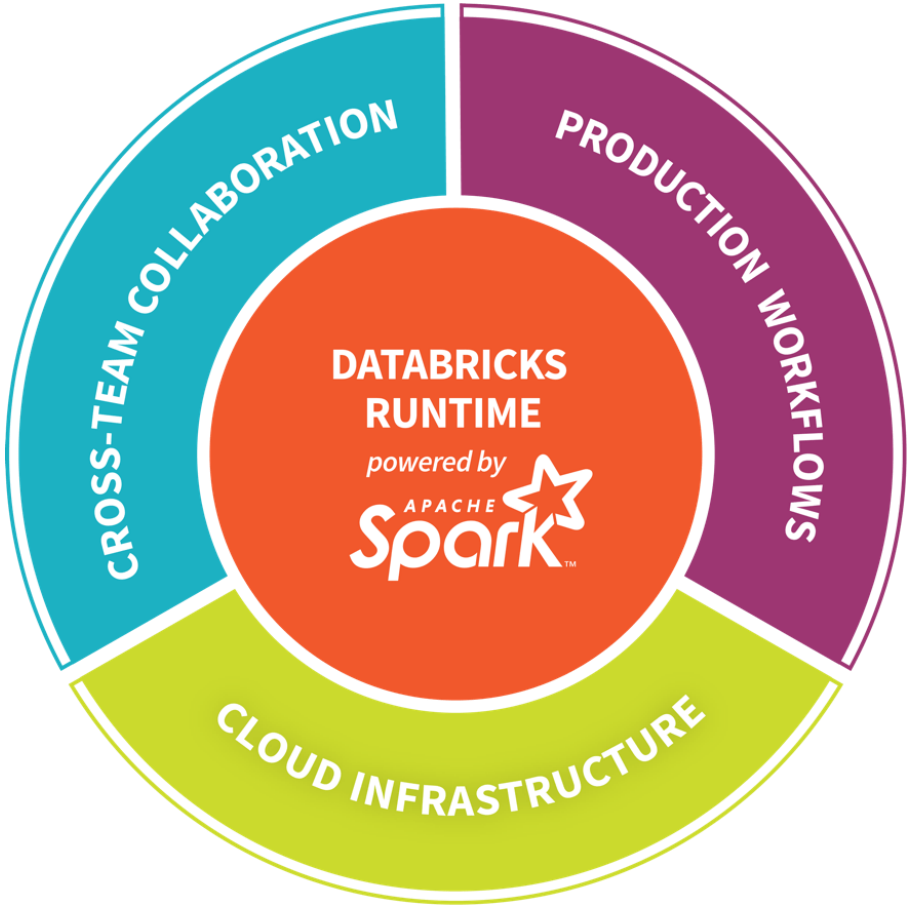
- In Azure



ORANGEMAN

DATABRICKS - COMPANY OVERVIEW

- Founded in late 2013
- By the creators of Apache Spark, original team from UC Berkeley AMPLab
- Largest code contributor code to Apache Spark
- Level 2/3 support partnership with
  - Hortonworks
  - MapR
  - DataStax
- Provides certifications such as Databricks Certified Application, Databricks Certified Distribution and Databricks Certified Developer
- Main Product: The Unified Analytics Platform
- In Oct 2017, introduced Databricks Delta (currently in private preview).



ORANGEMAN

Azure Databricks

- Azure Databricks is a first party service on Azure.
  - Unlike with other clouds, it is not an Azure Marketplace or a 3rd party hosted service.
- Azure Databricks is integrated seamlessly with Azure services:
  - Azure Portal: Service an be launched directly from Azure Portal
  - Azure Storage Services: Directly access data in Azure Blob Storage and Azure Data Lake Store
  - Azure Active Directory: For user authentication, eliminating the need to maintain two separate sets of uses in Databricks and Azure.
  - Azure SQL DW and Azure Cosmos DB: Enables you to combine structured and unstructured data for analytics
  - Apache Kafka for HDInsight: Enables you to use Kafka as a streaming data source or sink
  - Azure Billing: You get a single bill from Azure
  - Azure Power BI: For rich data visualization
- Eliminates need to create a separate account with Databricks.

Azure Databricks

AZURE DATA SOURCES

Data Lake Storage

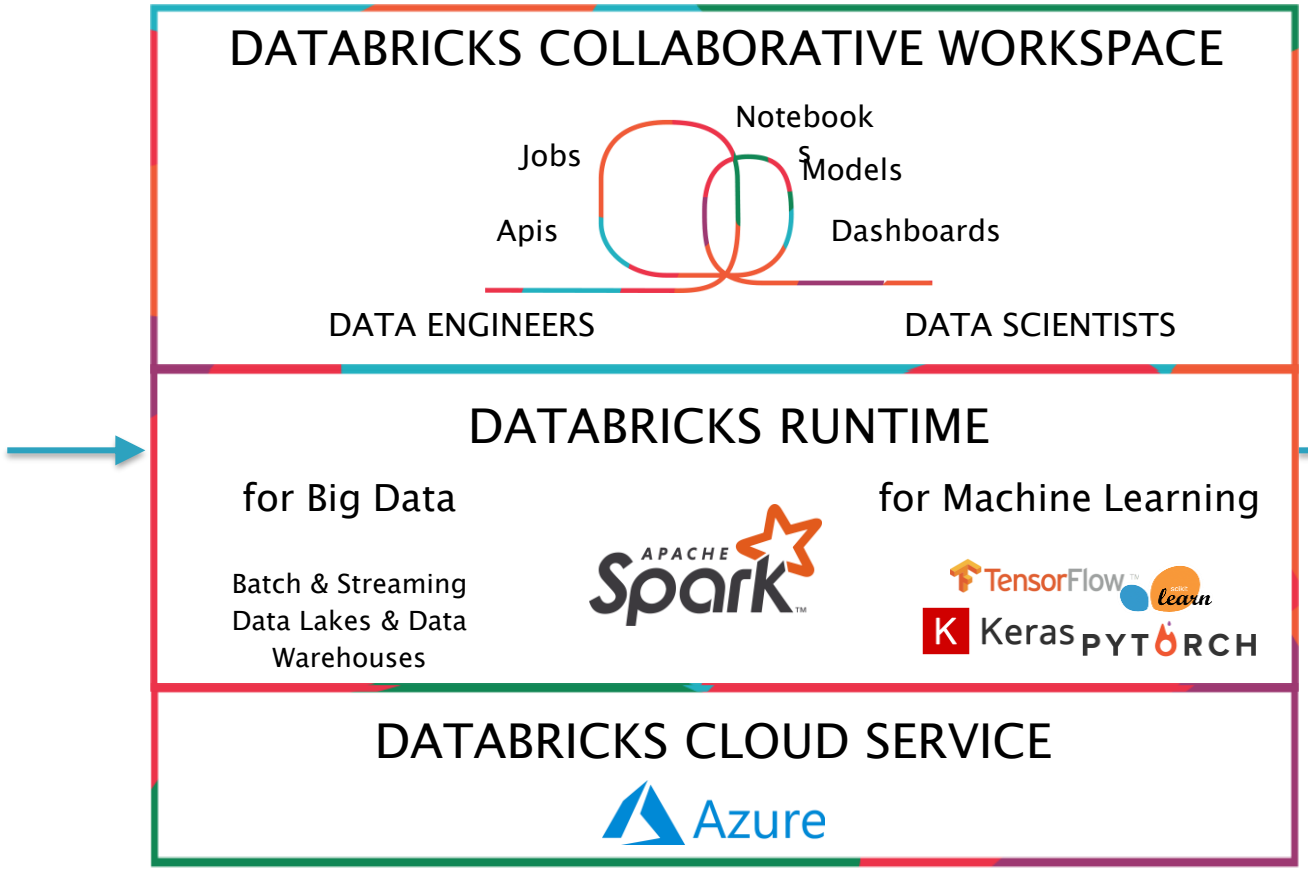
SQL Data Warehouse

SQL Databases

Cosmos DB

Event Hub

IoT Hub



BI Reporting
Dashboards

Microsoft Active Directory

Security Integration

Azure Portal

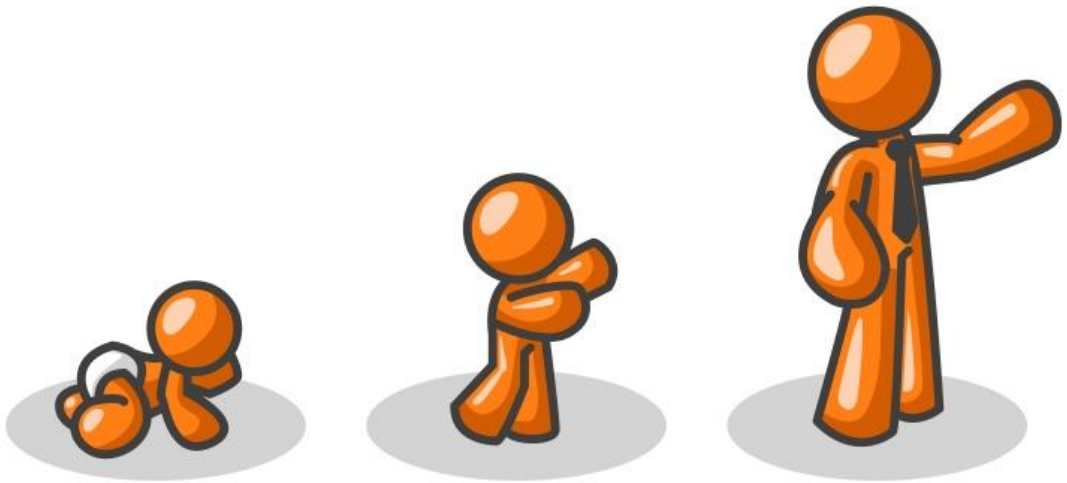
One-Click setup

Unified Billing

ORANGEMAN

Architecture

- Build on Apache Spark



ORANGEMAN

Why Spark?

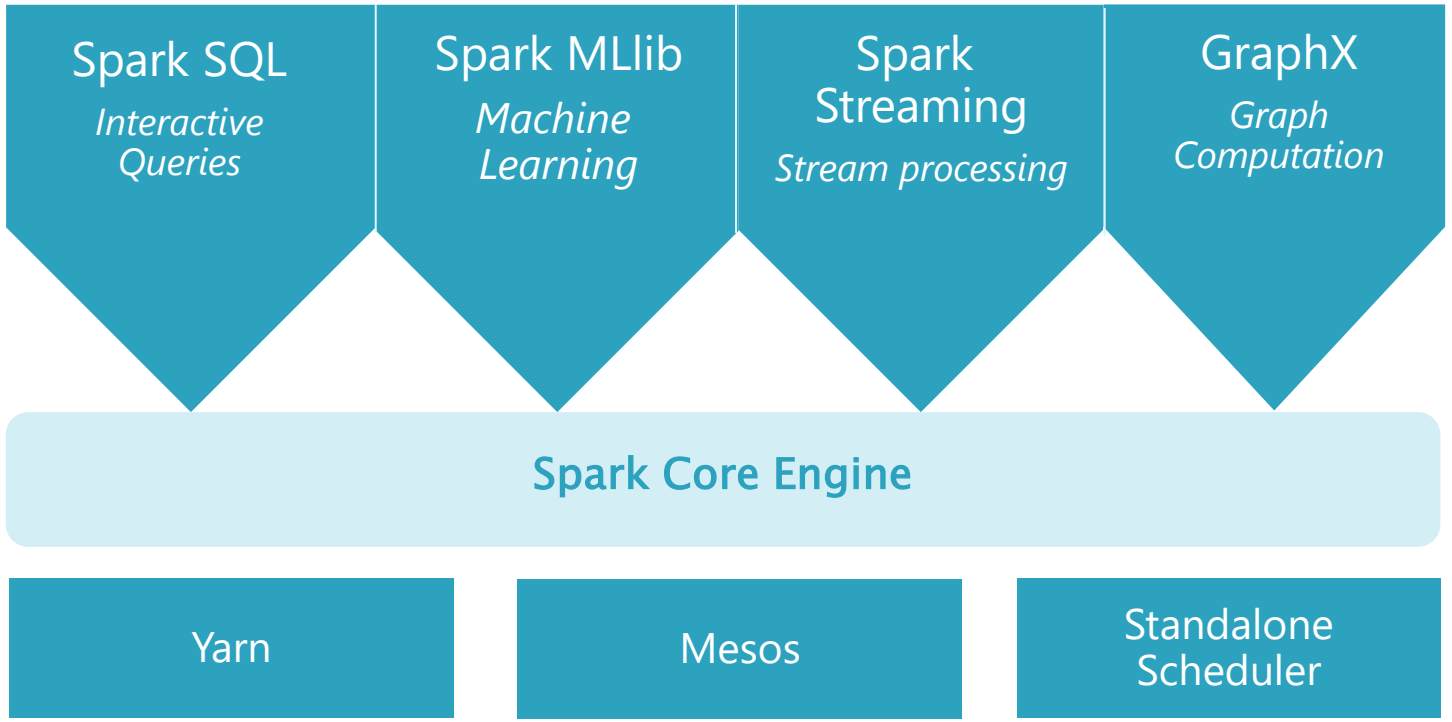
- Open-source data processing engine built around **speed, ease of use, and sophisticated analytics**
- In memory engine that is up to **100 times faster than Hadoop**
- **Largest open-source data project** with 1000+ contributors
- **Highly extensible** with support for Scala, Java and Python alongside Spark SQL, GraphX, Streaming and Machine Learning Library (Mllib)

ORANGEMAN

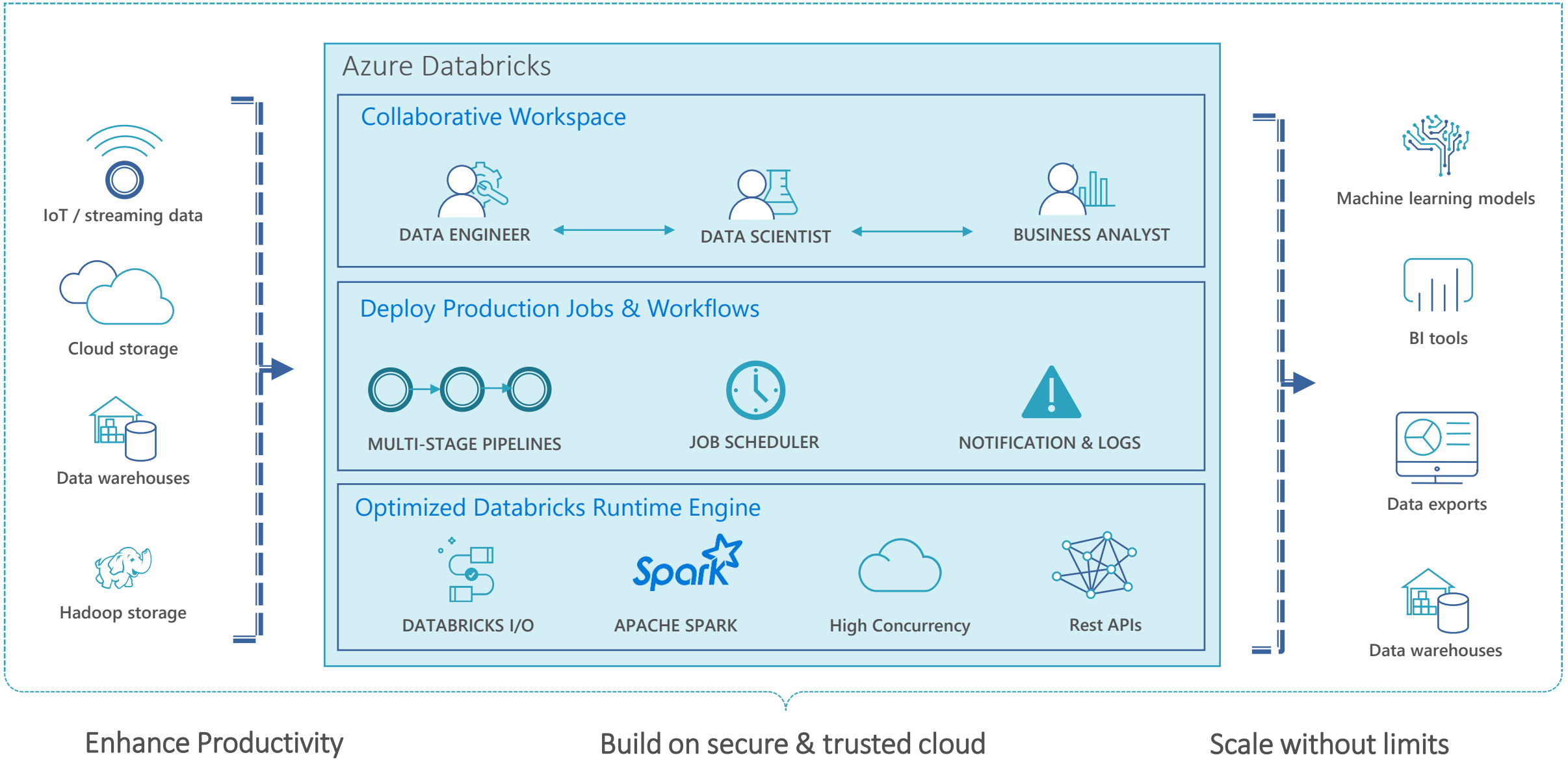
Apache Spark

An unified, open source, parallel, data processing framework for Big Data Analytics:

- Batch Processing
- Interactive SQL
- Real-time processing
- Machine Learning
- Deep Learning
- Graph Processing



ORANGEMAN



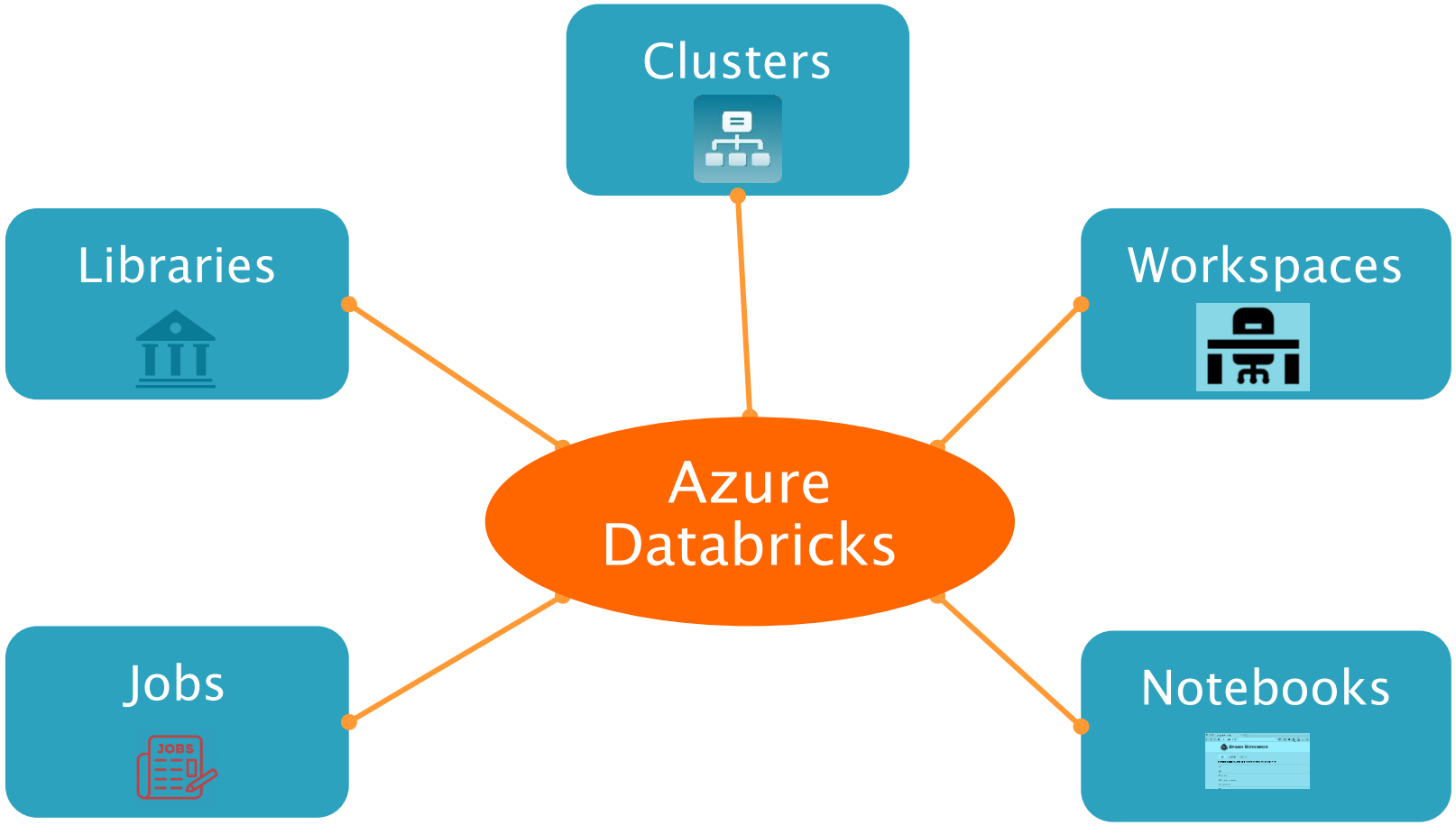
Enhance Productivity

Build on secure & trusted cloud

Scale without limits

ORANGEMAN

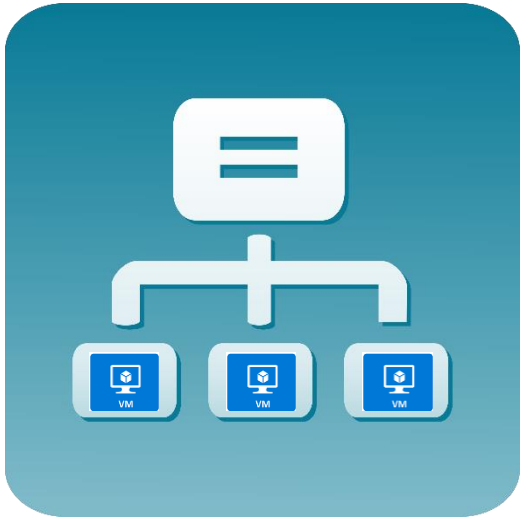
Azure Databricks Core Artifacts



ORANGEMAN

Clusters

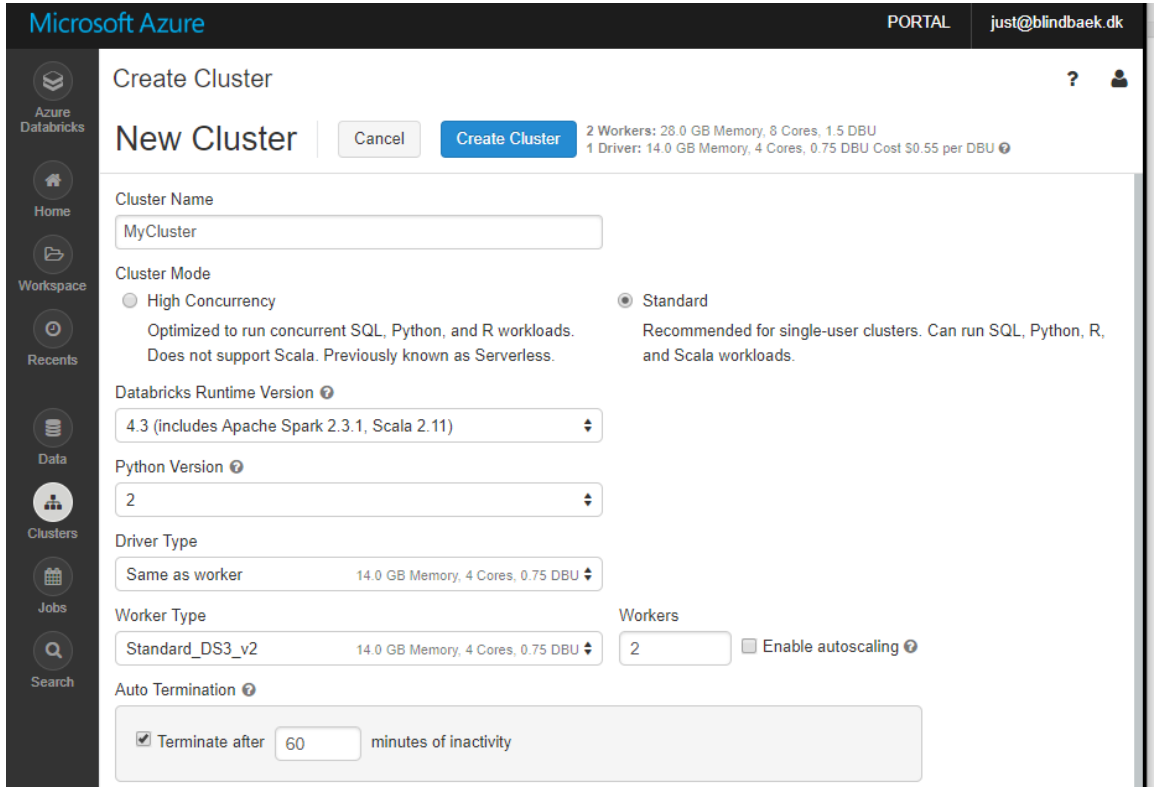
- Azure Databricks clusters are the set of Azure Linux VMs that host the Spark Worker and Driver Nodes
- Your Spark application code (i.e. Jobs) runs on the provisioned clusters.
- Azure Databricks clusters are launched in your subscription—but are managed through the Azure Databricks portal.
- Azure Databricks provides a comprehensive set of graphical wizards to manage the complete lifecycle of clusters—from creation to termination.

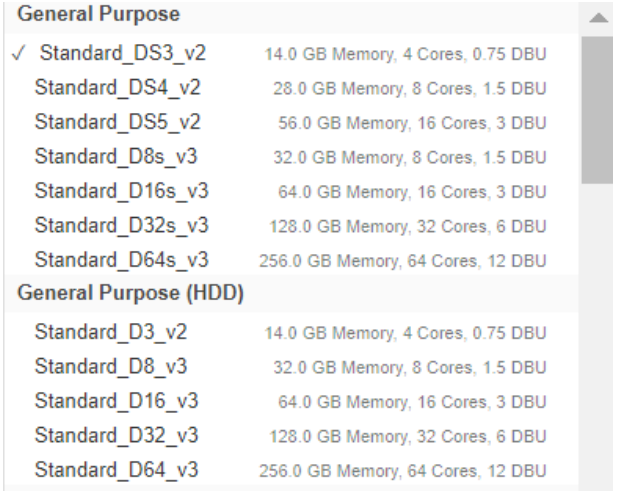


ORANGEMAN

Cluster creation

- You can create two types of clusters:
  - Standard
  - High Concurrency (Serverless)
- While creating a cluster you can specify:
  - Number of nodes
  - Autoscaling and Auto Termination policy
  - Auto Termination policy
  - Spark Configuration details
  - The Azure VM instance types for the Driver and Worker Nodes





ORANGEMAN

Demo

1. Create an Azure Databricks workspace
2. Login to Databricks
3. Create a Spark cluster in Databricks

Quickstart:

- https://docs.microsoft.com/en-us/azure/azure-databricks/quickstart-create-databricks-workspace-portal

ORANGEMAN

Azure Databricks Pricing

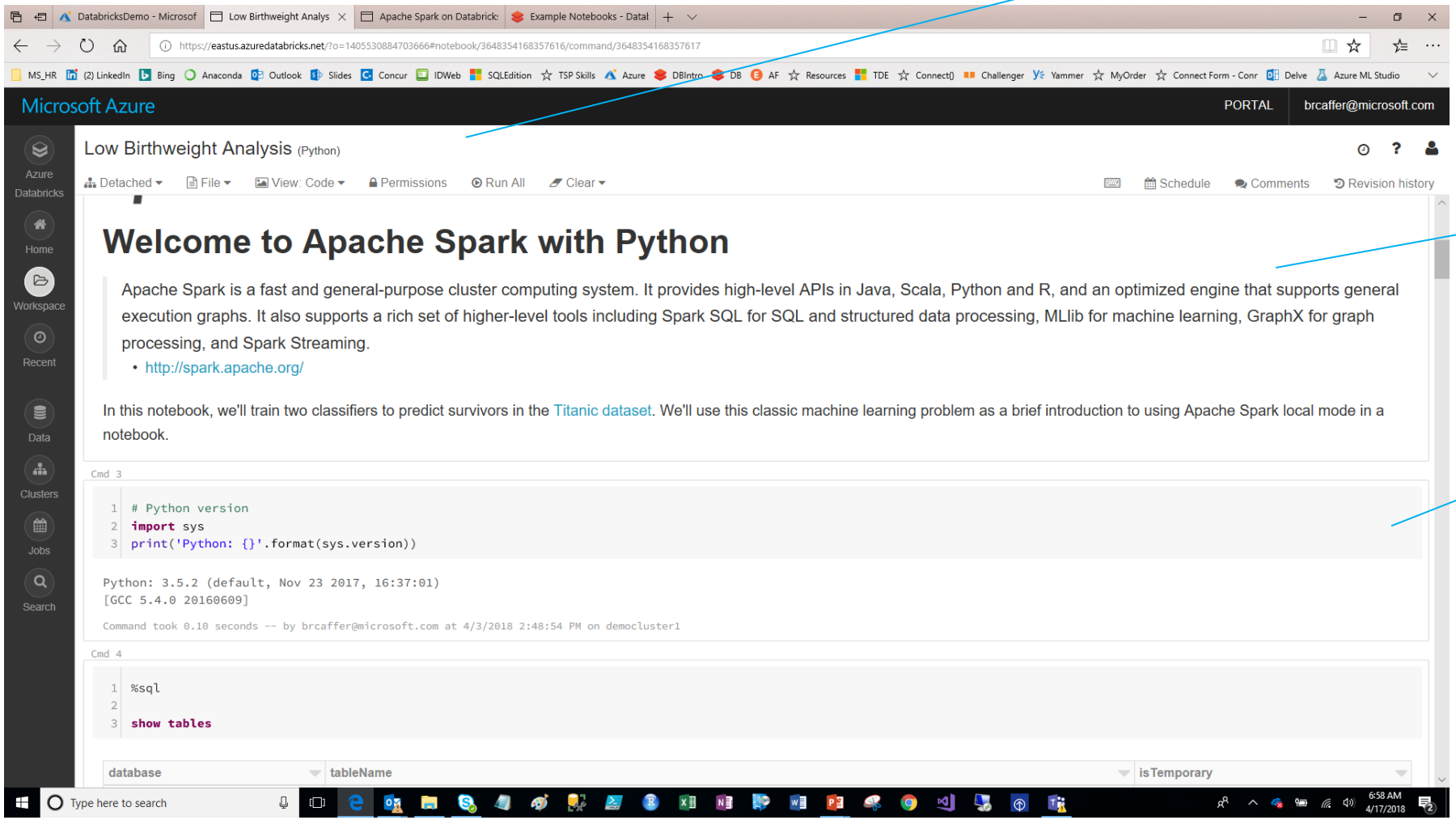
Workload	Standard Tier	Premium Tier
Data Engineering	$0.20/DBU-hour	$0.35/DBU-hour
Data Analytics	$0.40/DBU-hour	$0.55/DBU-hour

Instance	vCPU	RAM	DBU Count	Price	Price (3 year)
DS3 v2	4	14.00 GiB	0.75	$0.676/hour	$0.516/hour
DS4 v2	8	28.00 GiB	1.50	$1.352/hour	$1.031/hour
DS5 v2	16	56.00 GiB	3.00	$2.703/hour	$2.062/hour
E32 v3	32	256.00 GiB	8.00	$6.771/hour	$5.292/hour

- Instance prices are for Data Analytics Workload and Premium Tier
- Full list: https://azure.microsoft.com/en-us/pricing/details/databricks/

ORANGEMAN

Azure Databrick Notebooks



Toolbar

Annotations

Code Cells

ORANGEMAN

Demo

1. Download a sample data file
2. Create a notebook
3. Run a Spark SQL job

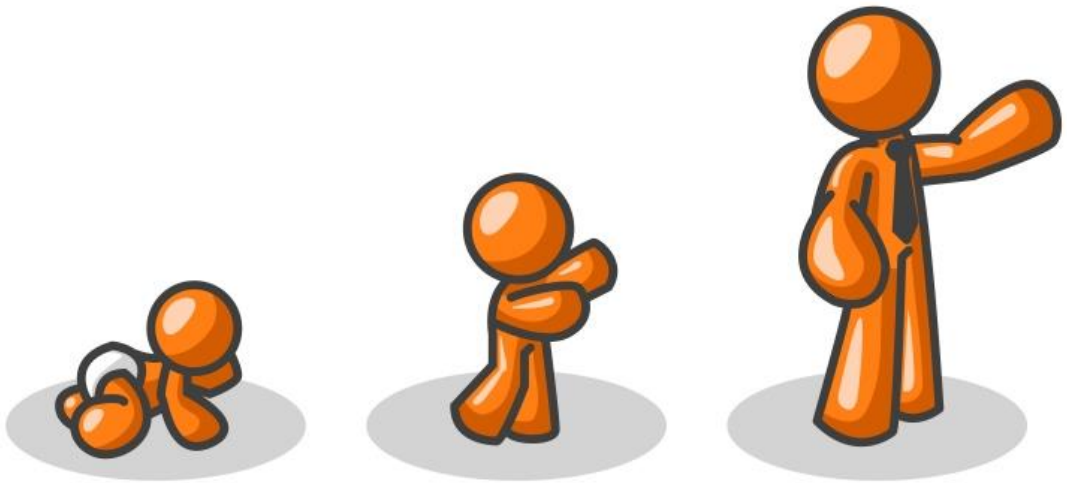
Quickstart:

- https://docs.microsoft.com/en-us/azure/azure-databricks/quickstart-create-databricks-workspace-portal

ORANGEMAN

Data processing

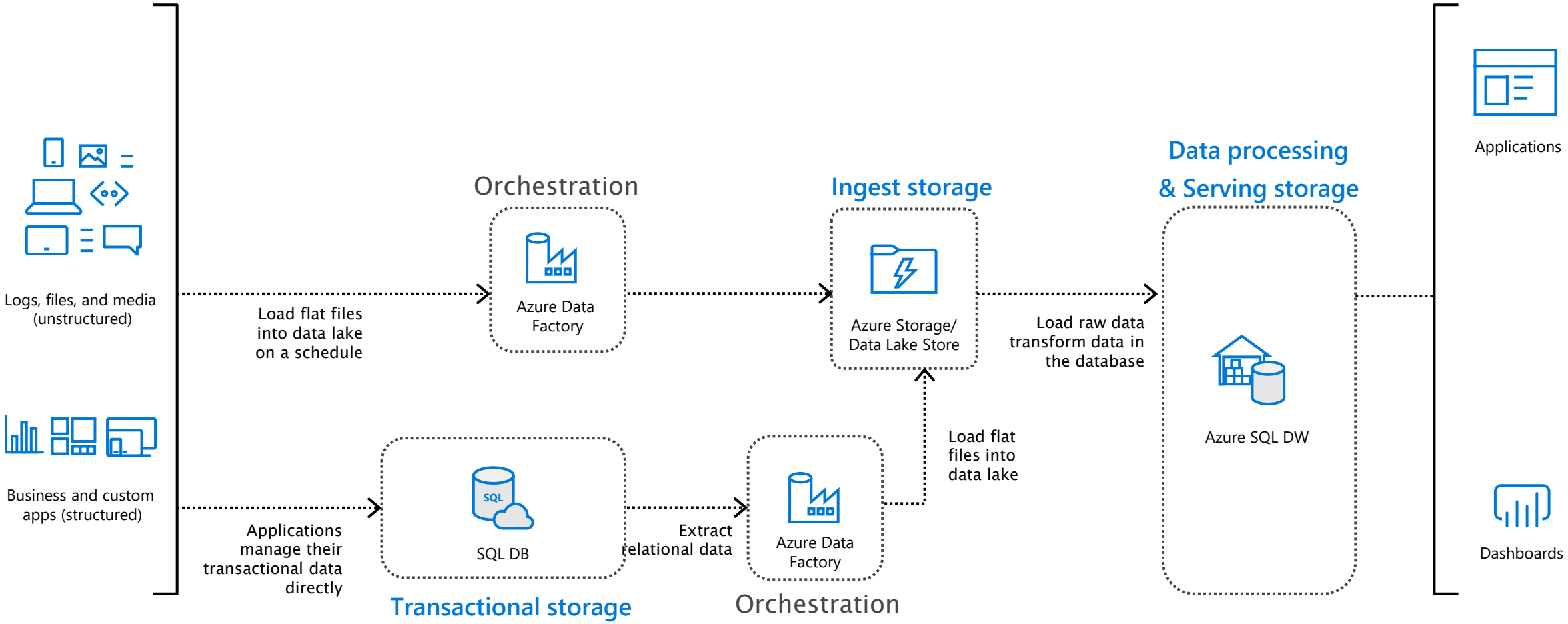
- with Azure Databricks



ORANGEMAN

Modern Data Warehouse Pattern #1

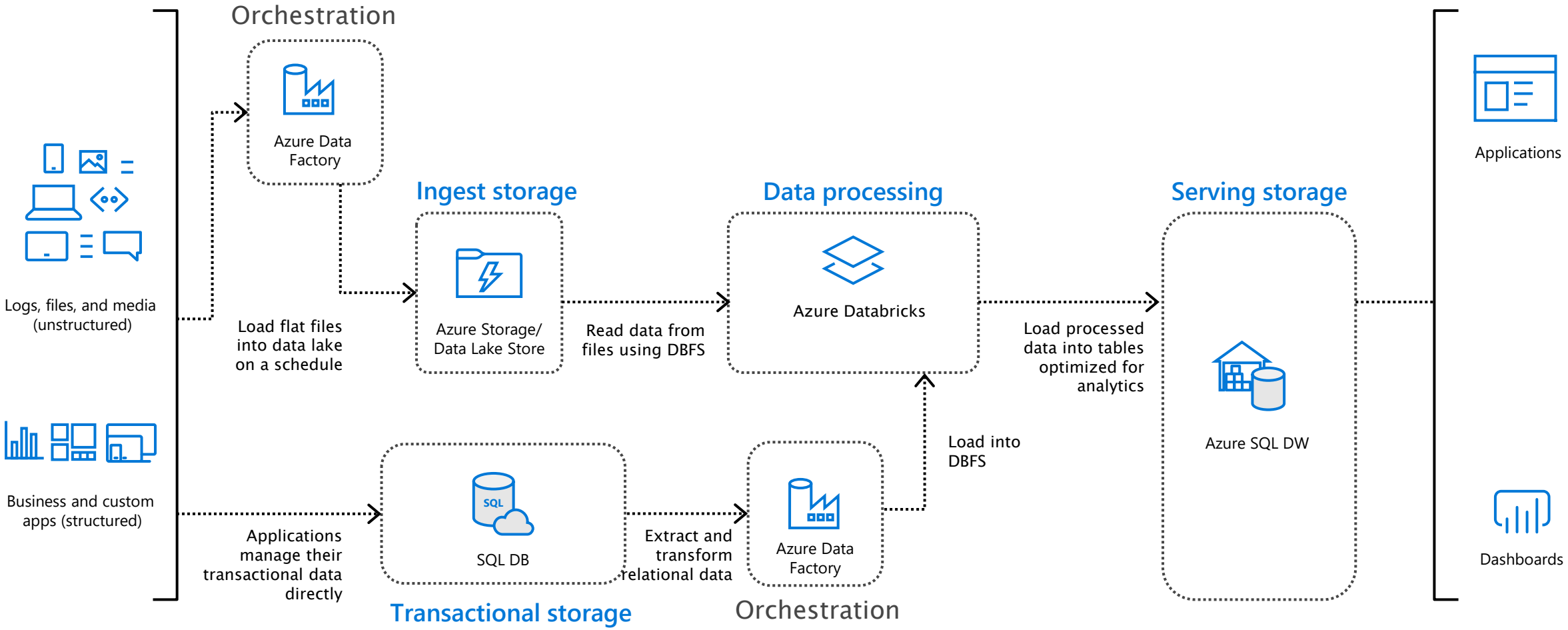
Data processing with Azure SQL Data Warehouse



© Microsoft Corporation

Modern Data Warehouse Pattern #2

Data processing with Azure Databricks



ORANGEMAN
© Microsoft Corporation

MODERN DATA WAREHOUSE

Azure also supports other Big Data services like Azure HDInsight and Azure Data Lake to allow customers to tailor the above architecture to meet their unique needs.

ORANGEMAN

Demo

1. Create an AAD service principal
2. Upload data to Data Lake Store
3. Associate service principal with Data Lake Store
4. Extract data from Data Lake Store
5. Transform data in Azure Databricks
6. Load data into Azure SQL Data Warehouse

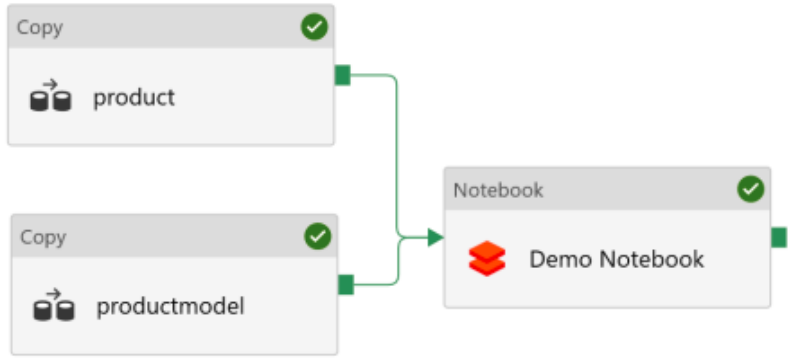
Quickstart:

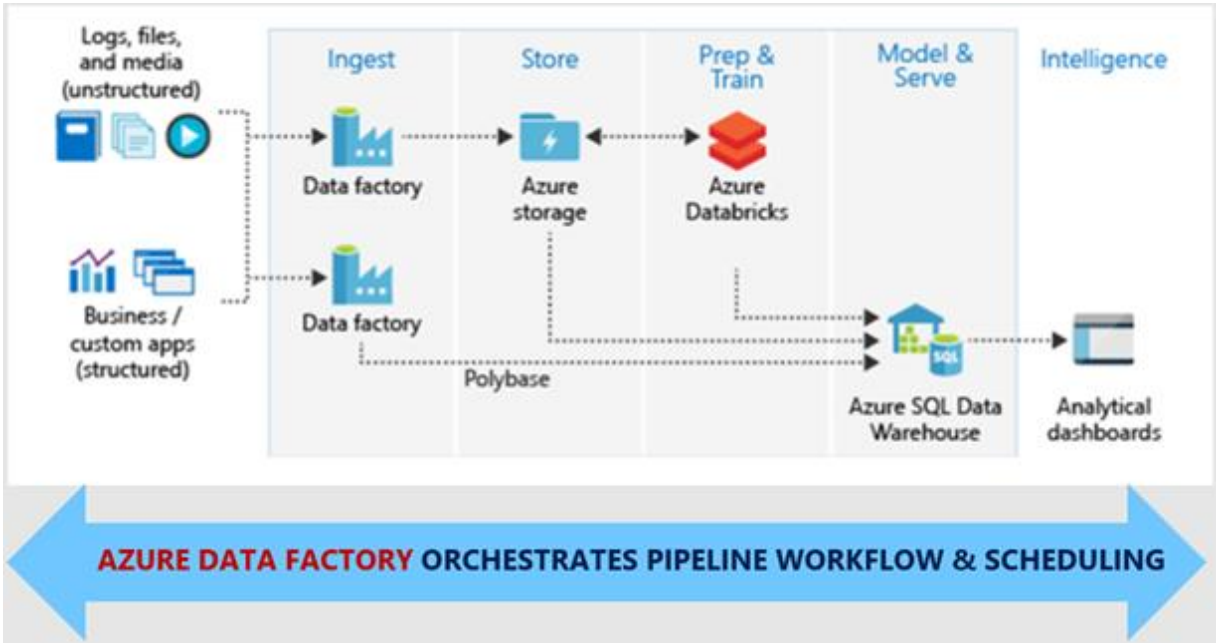
- https://docs.microsoft.com/en-us/azure/azure-databricks/databricks-extract-load-sql-data-warehouse

ORANGEMAN

Databricks notebook in Data factory

- Ingest data at scale using 70+ on-prem/cloud data sources
- Prepare and transform (clean, sort, merge, join, etc.) the ingested data in Azure Databricks as a Notebook activity step in data factory pipelines
- Monitor and manage your E2E workflow

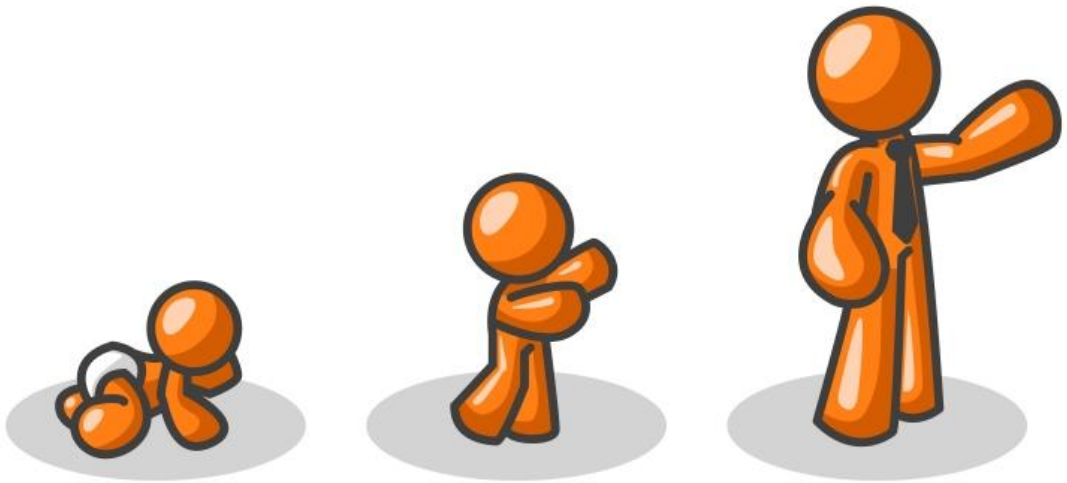




ORANGEMAN

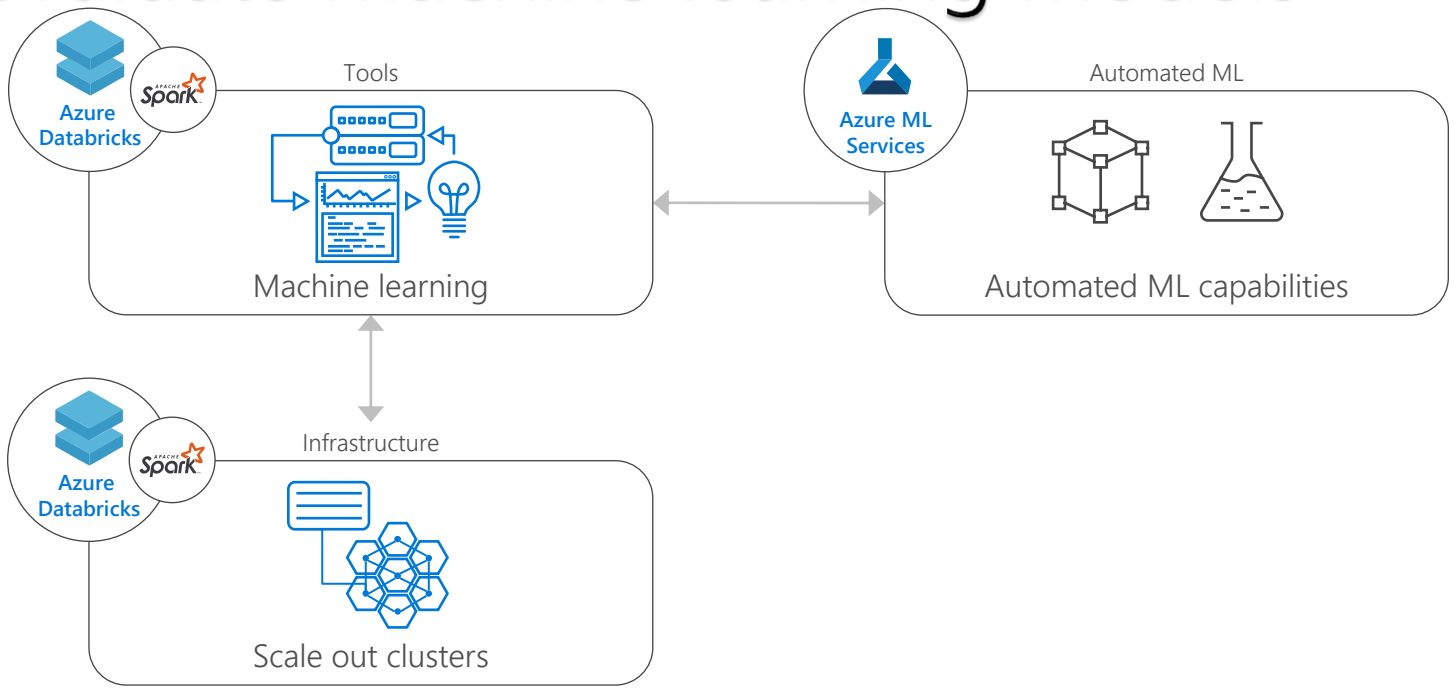
Machine Learning

- with Azure Databricks



ORANGEMAN

Train and evaluate machine learning models



Simplify model development

- Collaborate in interactive workspaces
- Access a library of battle-tested models
- Automate job execution

Scale compute resources to meet your needs

- Easily scale up or scale out
- Autoscale on serverless infrastructure
- Leverage commodity hardware

Quickly determine the right model for your data

- Determine the best algorithm
- Tune hyperparameters to optimize models
- Rapidly prototype in agile environments

ORANGEMAN

3 Ways for Machine Learning

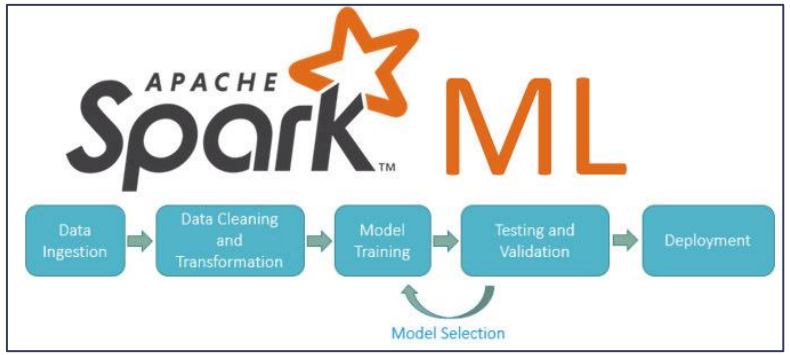
- #1 Scalable Machine Learning with Spark MLlib
  - Goal is to make practical machine learning extremely scalable and easy
  - Common Algorithms, Featurization, Pipelines, and Utilities need for ML
  - Subset of all ML techniques, but extremely scalable
- #2 Single Machine Data Science on Big Data with Azure Databricks
  - Use ADB to query “Big Data” stored on ADLS or Blob
  - Use Spark to Aggregate, Sample “Big Data” to make it “small data”
  - Collect this “small data” back to the driver for normal smaller data ML tools, R, Scikit-learn, etc
- #3 Scale Out / Parallelization for Single Machine Data Science
  - Combination of the above two
  - Use Databricks for cross validation, training a bunch of small models, etc
  - Apply user defined functions from R and Python

ORANGEMAN

SPARK MACHINE LEARNING(ML) OVERVIEW

Enables Parallel, Distributed ML for large datasets on Spark Clusters

- Offers a set of parallelized machine learning algorithms (see next slide)
- Supports Model Selection (hyperparameter tuning) using Cross Validation and Train-Validation Split.
- Supports Java, Scala or Python apps using DataFrame-based API (as of Spark 2.0). Benefits include:
  - An uniform API across ML algorithms and across multiple languages
  - Facilitates ML pipelines (enables combining multiple algorithms into a single pipeline).
  - Optimizations through Tungsten and Catalyst
- Spark MLlib comes pre-installed on Azure Databricks
- 3rd Party libraries supported include: H20 Sparkling Water, SciKit-learn and XGBoost



ORANGEMAN

SPARK ML ALGORITHMS

Spark ML Algorithms

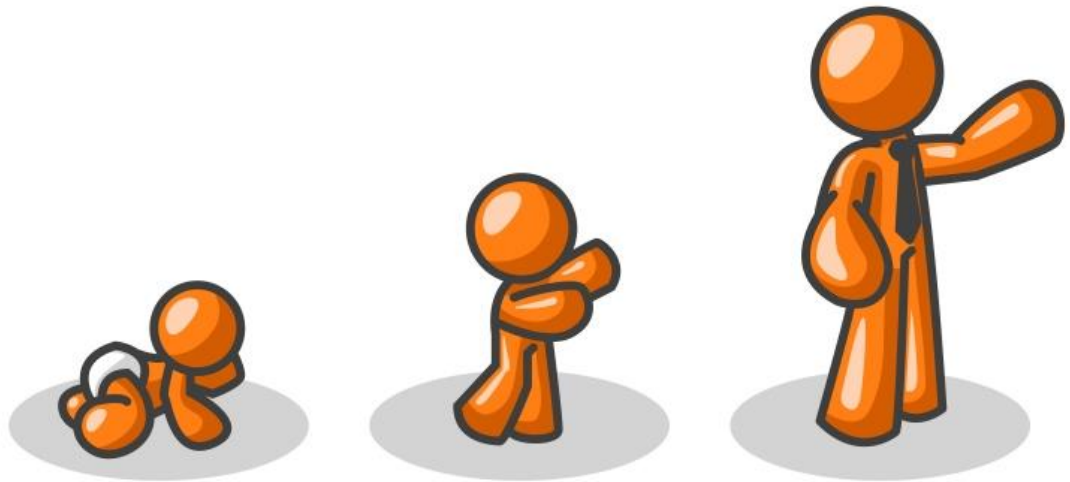
Classification and Regression	• Linear Models (SVMs, logistic regression, linear regression) • Naïve Bayes • Decision Trees • Ensembles of trees (Random Forest, Gradient-Boosted Trees) • Isotonic regression
Clustering	• k-means and streaming k-means • Gaussian mixture • Power iteration clustering (PIC) • Latent Dirichlet allocation (LDA)
Collaborative Filtering	• Alternating least squares (ALS)
Dimensionality Reduction	• SVD • PCA
Frequent Pattern Mining	• FP-growth • Association rules
Basic Statistics	• Summary statistics • Correlations • Stratified sampling • Hypothesis testing • Random data generation

ORANGEMAN

Why use Azure Databricks for Machine learning?

- Complete platform in one (Data ingestion, exploration, transformation, featurization, model building, model tuning, and even model serving).
- No need to copy the data in our system to do ML on it.
- Data Scientists like the ease of use of our platform.
- Deep learning algorithms are now available!
- Productionization Features built in.

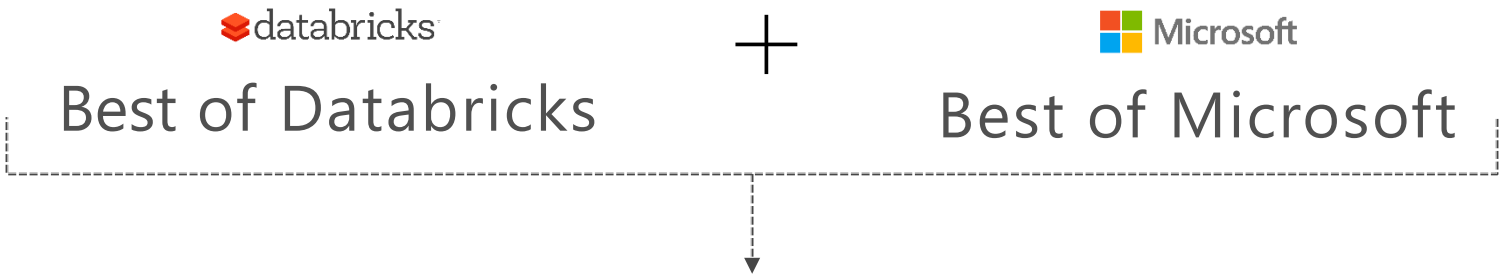
Azure Databricks recap



ORANGEMAN

What is Azure Databricks?

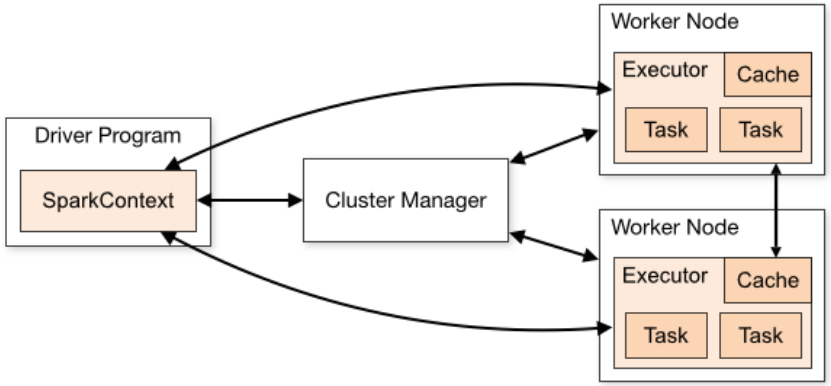
A fast, easy and collaborative Apache® Spark™ based analytics platform optimized for Azure

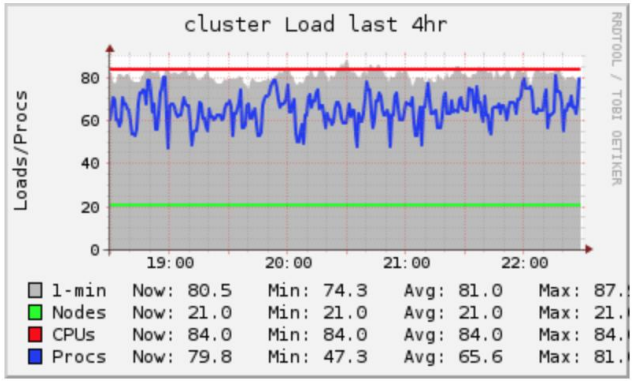


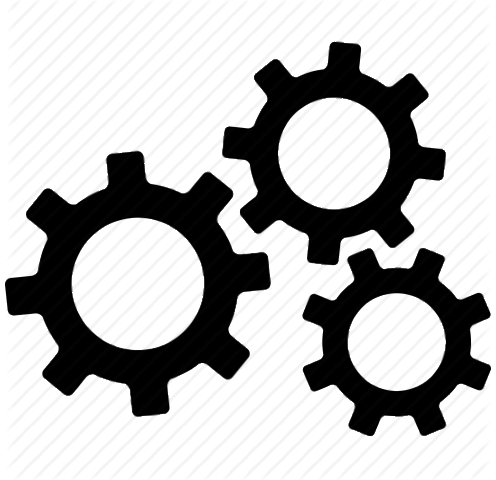
- Designed in collaboration with the founders of Apache Spark
- One-click set up; streamlined workflows
- Interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.
- Native integration with Azure services (Power BI, SQL DW, Cosmos DB, ADLS, Azure Storage, Azure Data Factory, Azure AD, Event Hub, IoT Hub, HDInsight Kafka, SQL DB)
- Enterprise-grade Azure security (Active Directory integration, compliance, enterprise-grade SLAs)

ORANGEMAN

Infinite Scale, Lower Cost, Zero Management







1 to 1000s of Worker Nodes

Auto-scale Compute & Storage

Auto-Recovery & Upgrade

ORANGEMAN

DATABRICKS ACCESS CONTROL

Access control can be defined at the user level via the Admin Console

Access Control can be defined for Workspaces, Clusters, Jobs and REST APIs

Databricks Access Control

Workspace Access Control	Defines who can who can view, edit, and run notebooks in their workspace
Cluster Access Control	Allows users to who can attach to, restart, and manage (resize/delete) clusters. Allows Admins to specify which users have permissions to create clusters
Jobs Access Control	Allows owners of a job to control who can view job results or manage runs of a job (run now/cancel)
REST API Tokens	Allows users to use personal access tokens instead of passwords to access the Databricks REST API

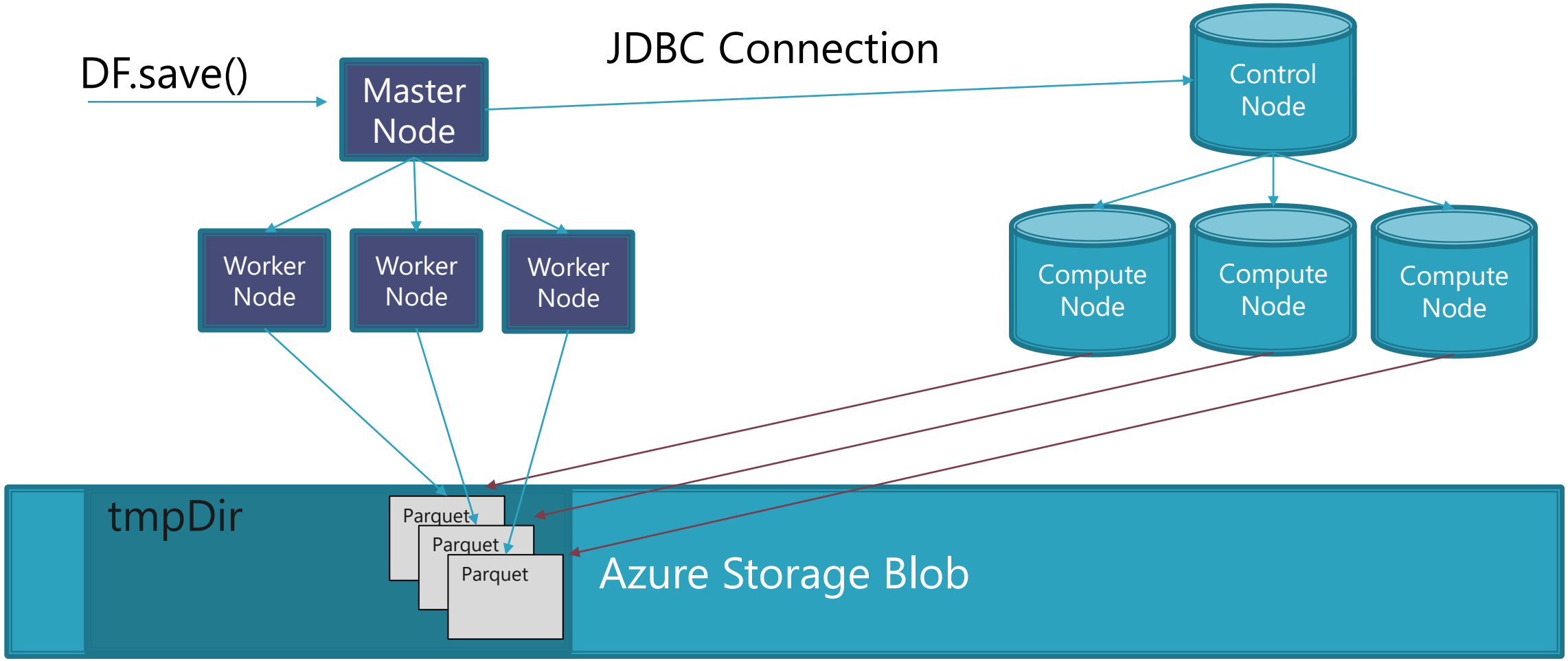
ORANGEMAN

Your Language, Your Data (Anywhere), Your Format

- SQL, Python, Scala & R Support
  - Code in your favorite language
- Source data from File System, Object stores, HDFS, Database, Pub-Sub systems & Others
  - Read and write data from/to multiple sources
  - Optimized for Azure Blob Store, ADLS, SQLDW, Event Hubs & Cosmos DB
- File Formats
  - CSV, JSON, Parquet, Text, ORC, XML & More

ORANGEMAN

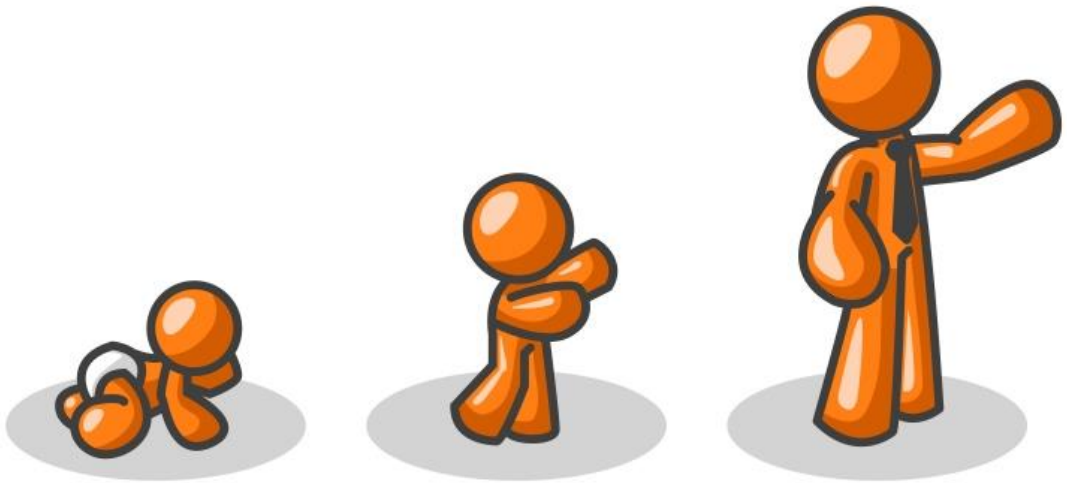
SQL Data Warehouse Connector



ORANGEMAN

Azure Databricks Delta

- Introducing



ORANGEMAN

Why Databricks Delta?

2. Spark Query Performance

Fast at Scale (10-100x Faster)
Cheaper to Operate
Indexing, Statistics & Caching

1. Data Reliability

ACID Compliant Transactions
Schema Enforcement & Evolution

LOTS OF NEW DATA

User Behavior Data

Click Streams

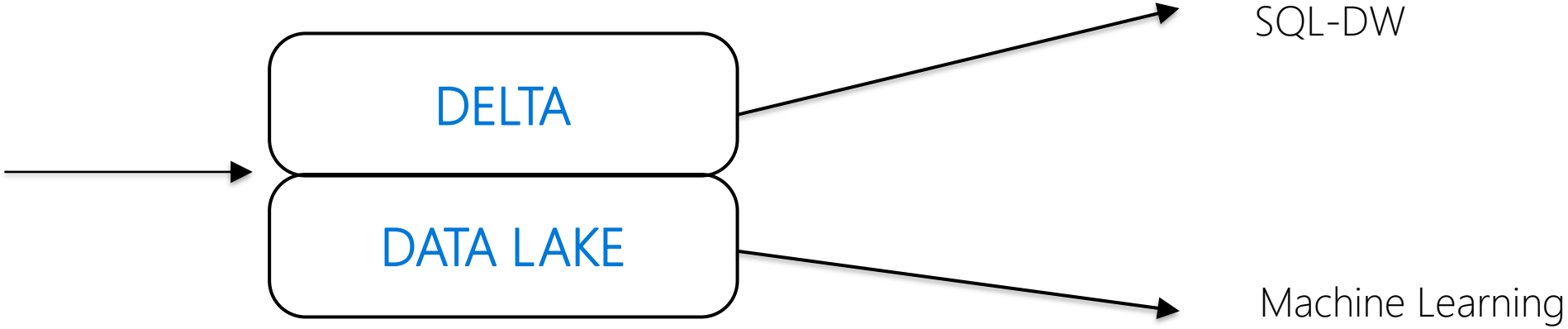
Sensor data (IoT)

Video/Speech

Usage/Billing data

Machine Telemetry

Commerce Data



3. Simplified Architecture

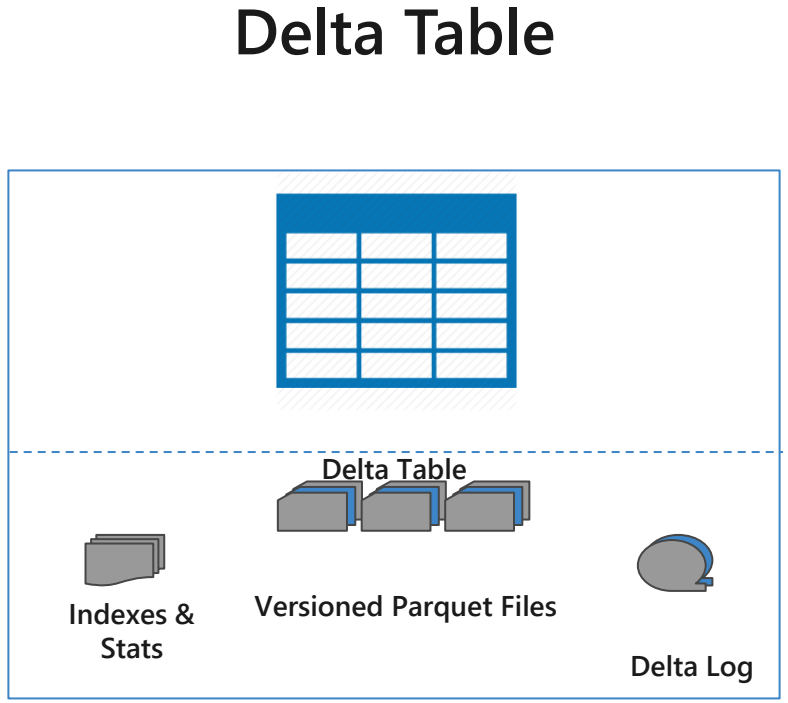
Unify batch & streaming
Early data availability for analytics

ORANGEMAN

Azure Databricks Delta Architecture

Delta Table = Parquet + Transaction Log

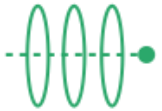
- Linear history of atomic changes
- Optimistic Concurrency Control
- Log checkpoint is stored as Parquet
- Lazy GC = Free Snapshot Isolation



ORANGEMAN

Azure Databricks Delta

Handle terabytes & petabytes of data



Low latency streaming ingestion

Avoid corrupt & messy data while reading & writing

Control on how to adapt to changing schema

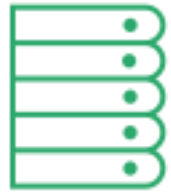
Enable scientists & analysts to read data quickly for interactive analysis - Indexing

ORANGEMAN

Azure Databricks Delta – Fast Reads

Data format

Compaction



Partitioning

Indexing

Caching

ORANGEMAN