Microsoft Partner of the Year
2015 Winner
Big Data and Analytics

BIG DATA PROCESSING

# A DEEP DIVE IN HADOOP/SPARK & AZURE SQL DW

Presented By: Orion Gebremedhin
Director of Technology, Data & Analytics,
Neudesic LLC.
Data Platform VTSP, Microsoft Corp.
@OrionGM

TOPICS

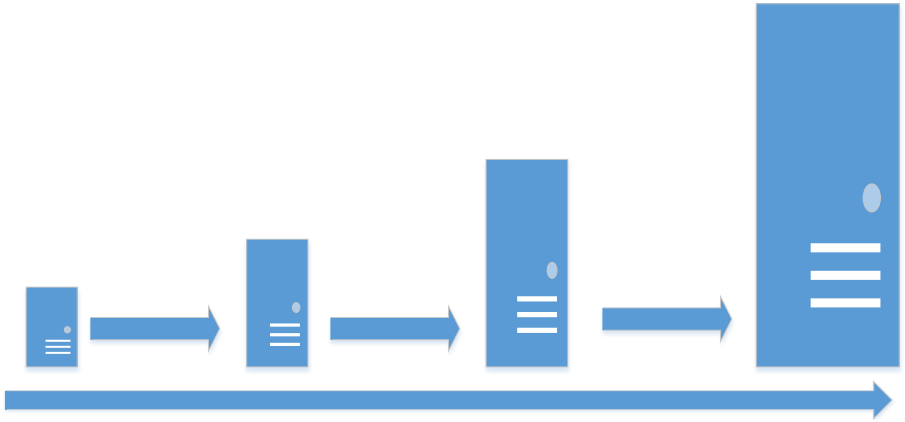
# COVERED

1. Fundamentals of Big Data Platforms
2. Major Big Data Tools

Scaling

# Up vs. Out

## SCALE UP (SMP)



Upgrade components or buy bigger server each time

Multiprocessor system where processors share resources :

- Operating System (OS)
- Memory
- I/O devices connected using a common bus

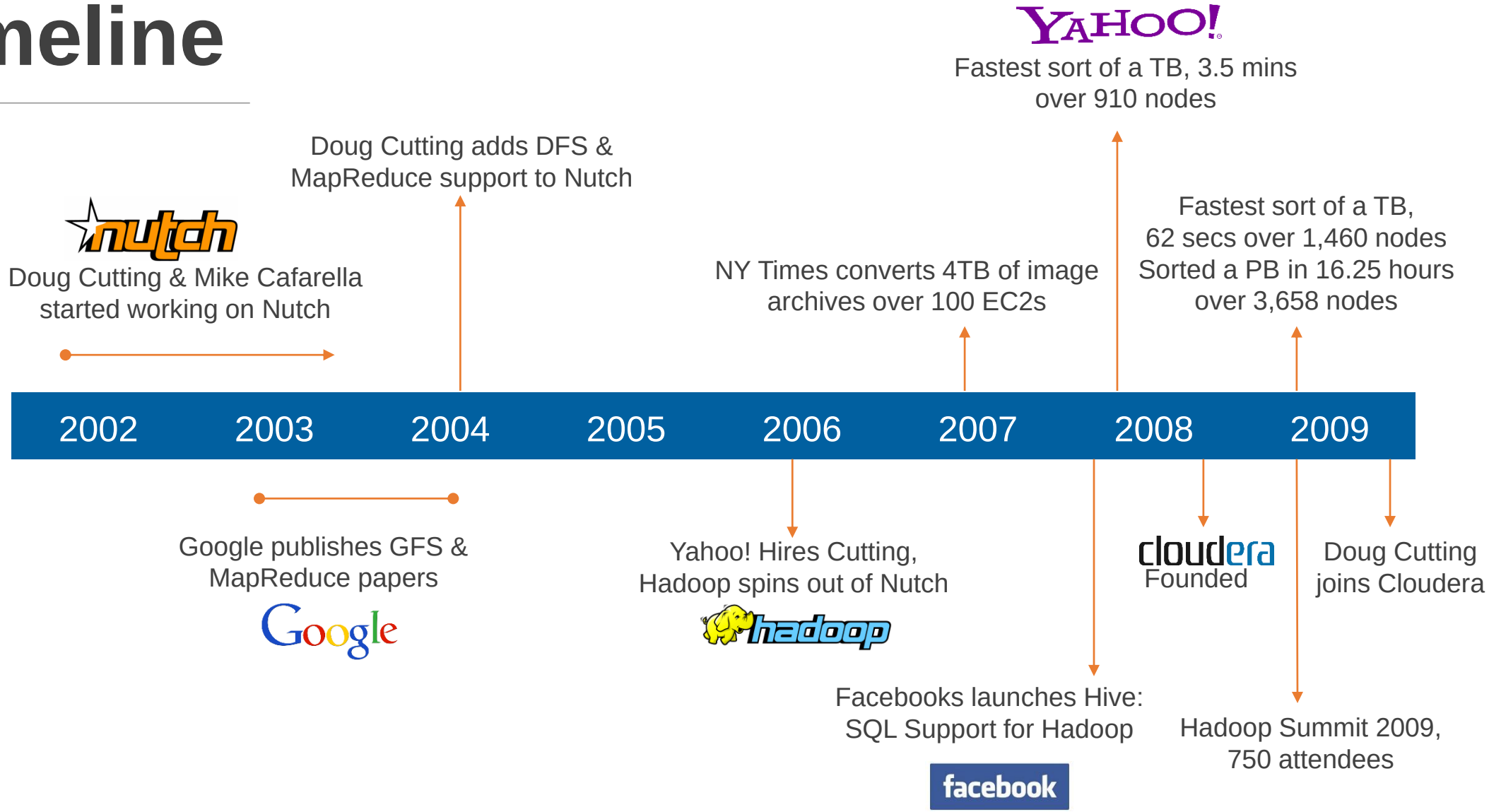
## SCALE OUT (MPP)


Add nodes to the cluster

- Multiple processing nodes
- OS
- RAM
- Network

Innovation

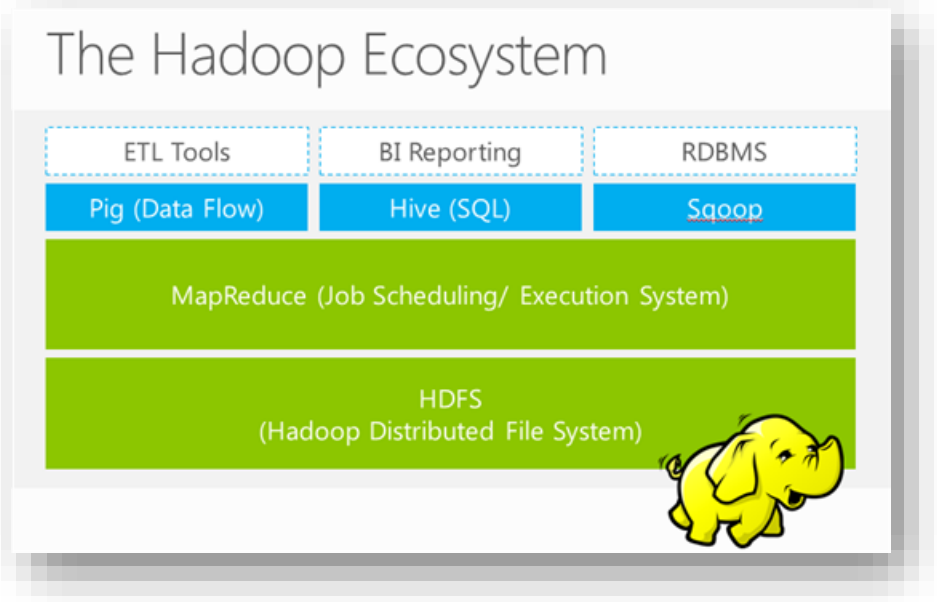
# Timeline



THE

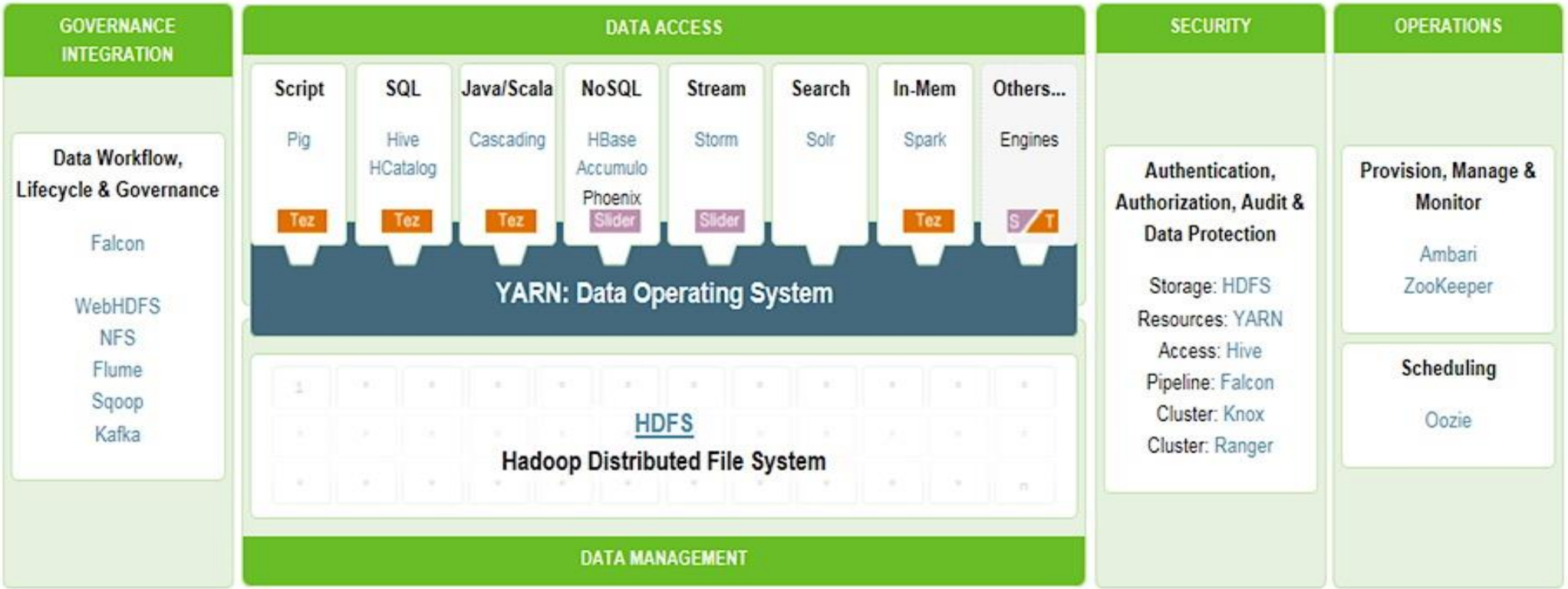
# FUNDAMENTALS OF HADOOP

- Hadoop evolved directly from commodity scientific supercomputing clusters developed in the 1990s
- Hadoop consists of:
  - MapReduce
  - Hadoop Distributed File System (HDFS)



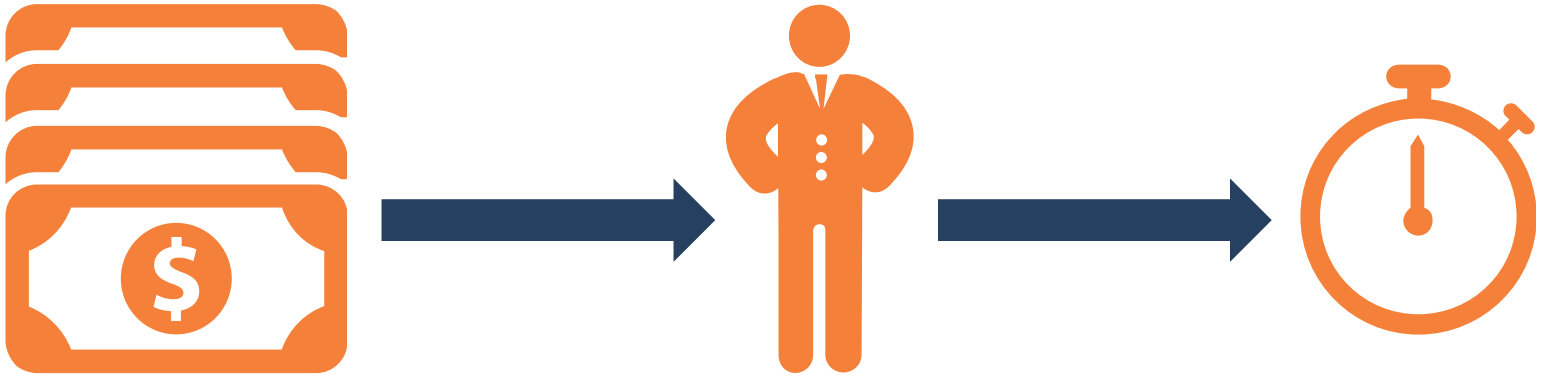
WHAT'S

# NEW…



BASICS OF

# MPP



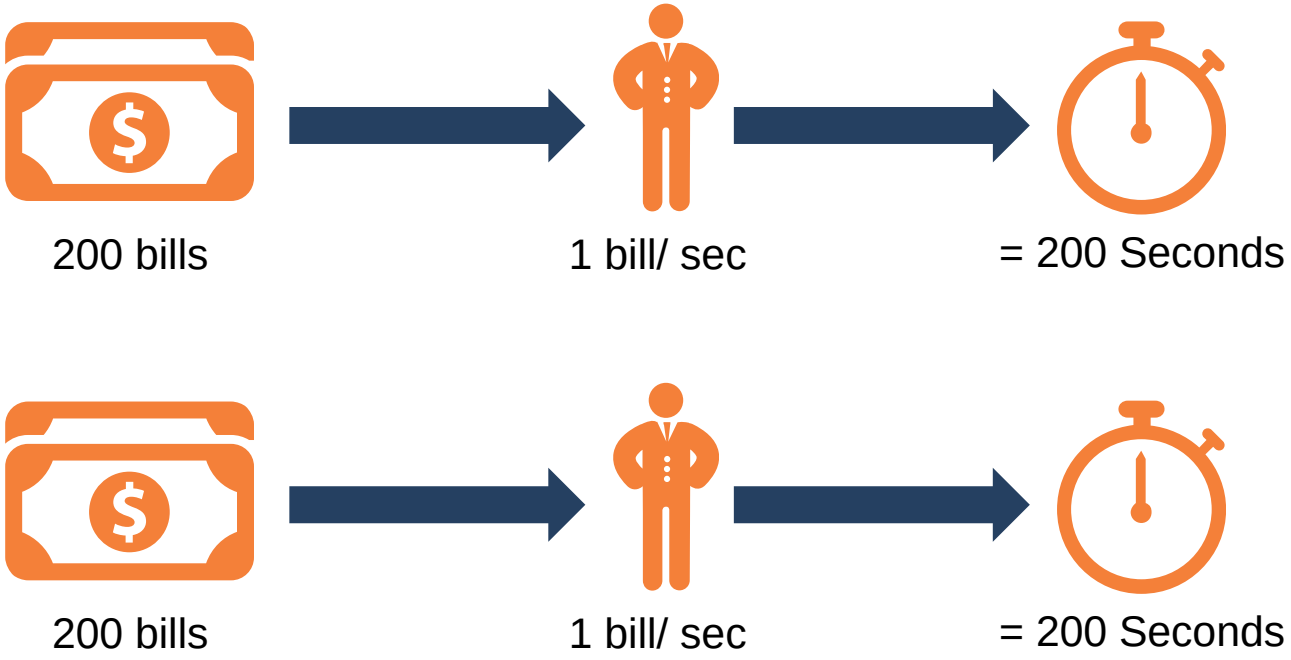
400 bills

1 bill/ sec

= 400 Seconds

BASICS OF

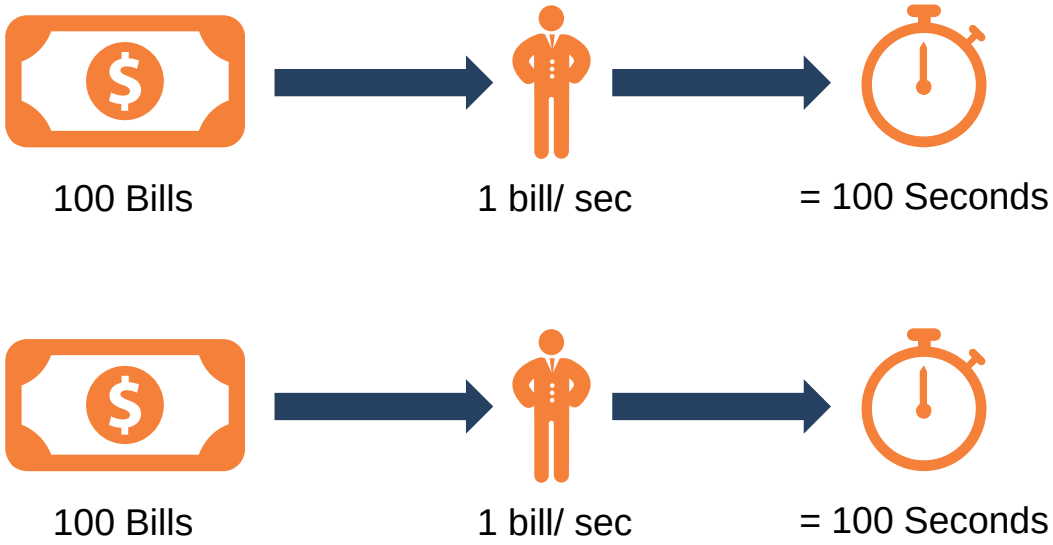
# MPP

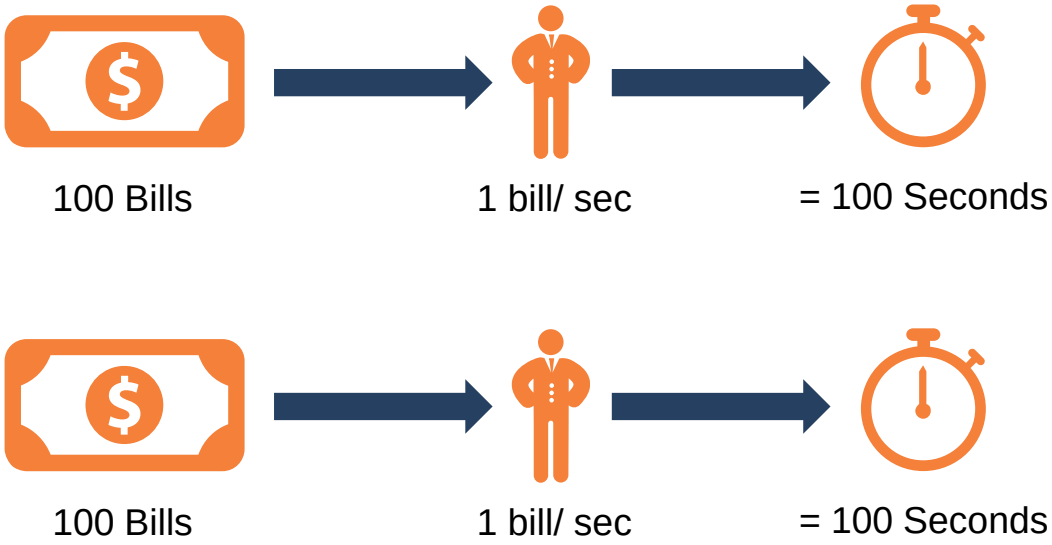



Total = 200 Seconds

BASICS OF

# MPP



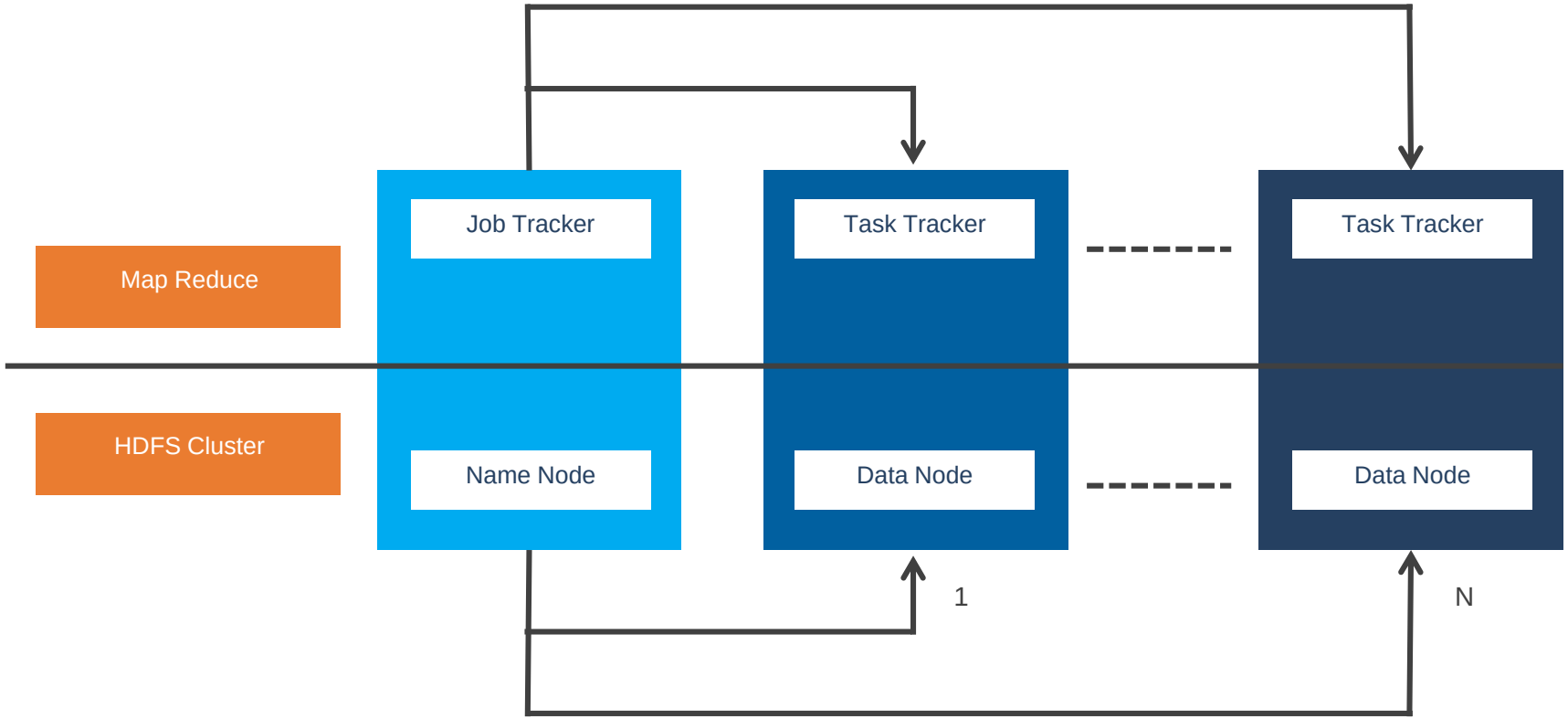




Total = 100 Seconds

HDFS &

# MAPREDUCE

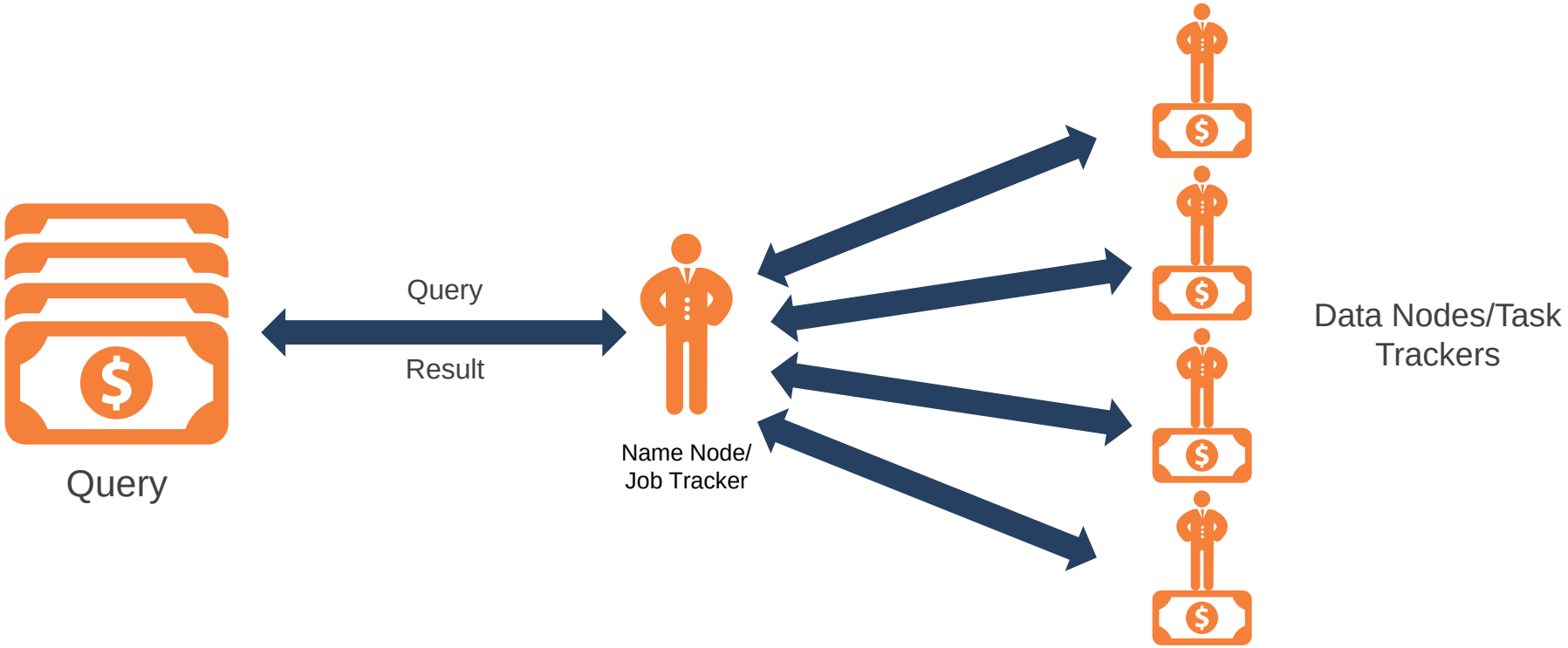
*The Main Node: runs the Job tracker and the name node controls the files.*
*Each node runs two processes: Task Tracker and Data Node*



BASICS OF

# MAPREDUCE

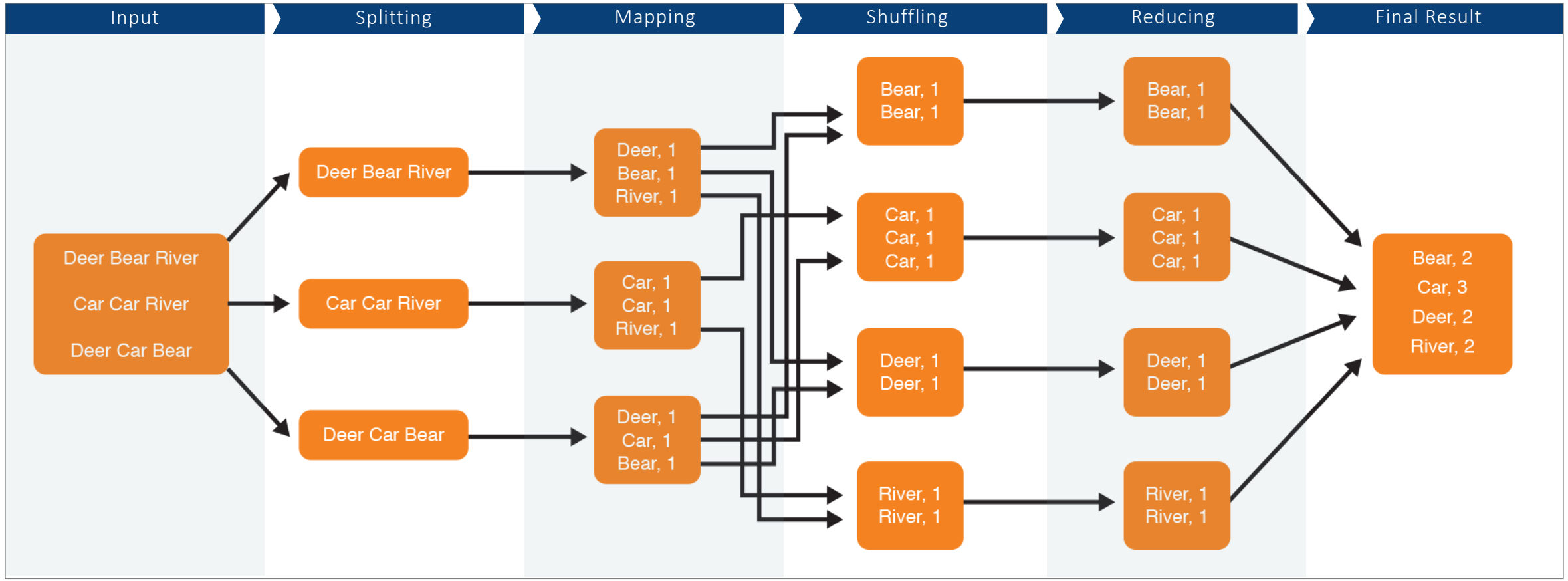
*The Main Node: runs the Job tracker and the name node controls the files.*
*Each node runs two processes: Task Tracker and Data Node*



EXECUTION UNITS

# MAPREDUCE

The overall MapReduce word count process



SOME DISTRIBUTIONS OF

# APACHE HADOOP

Apache Foundation

hadoop

cloudera

Hortonworks

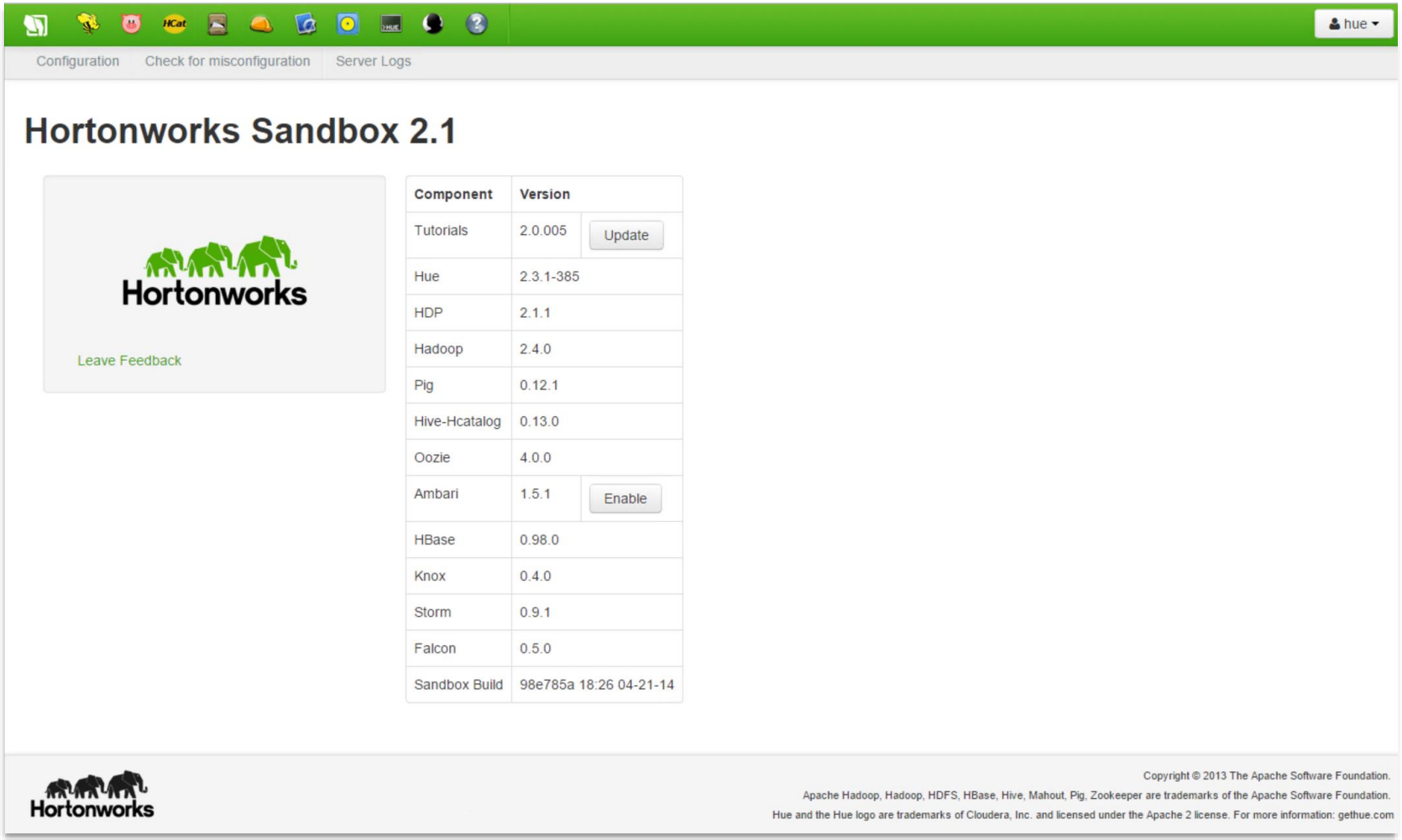
amazon web services™

pentaho

MAPR

Sandbox

# Hortonworks



MAPREDUCE

# PIG & HIVE

## MAPREDUCE

- Java
- Write many lines of code

## PIG

- Mostly used by Yahoo
- Most used for data processing
- Shares some constructs w/ SQL
- Is more Verbose
- Needs a lot of training for users with limited procedural programming background
- Offers control over the flow of data

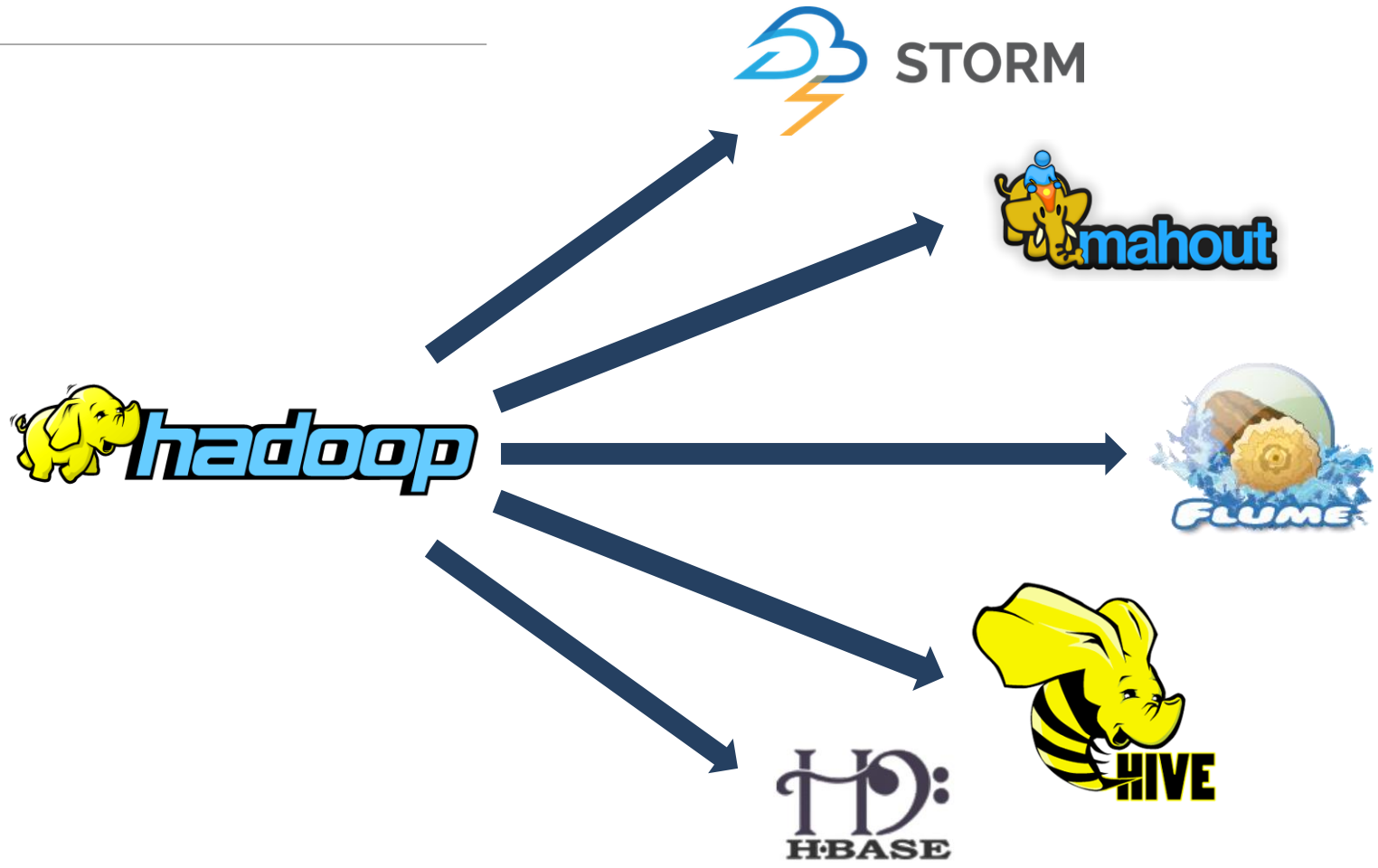
## HIVE

- Mostly used by Facebook for analytic purposes
- Used for analytics
- Relatively easier for developers w/ SQL experience
- Less control over optimization of data flows compared to Pig

*Not as efficient as MapReduce*
*Higher productivity for data scientists and developers*

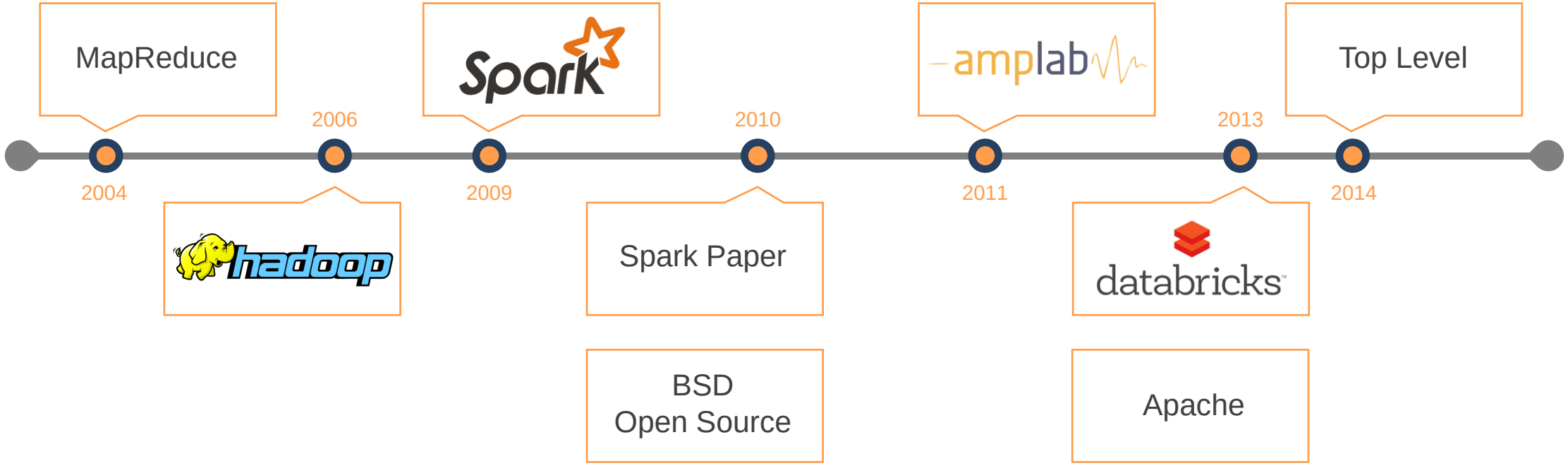
THE

# EXPLOSION OF HADOOP



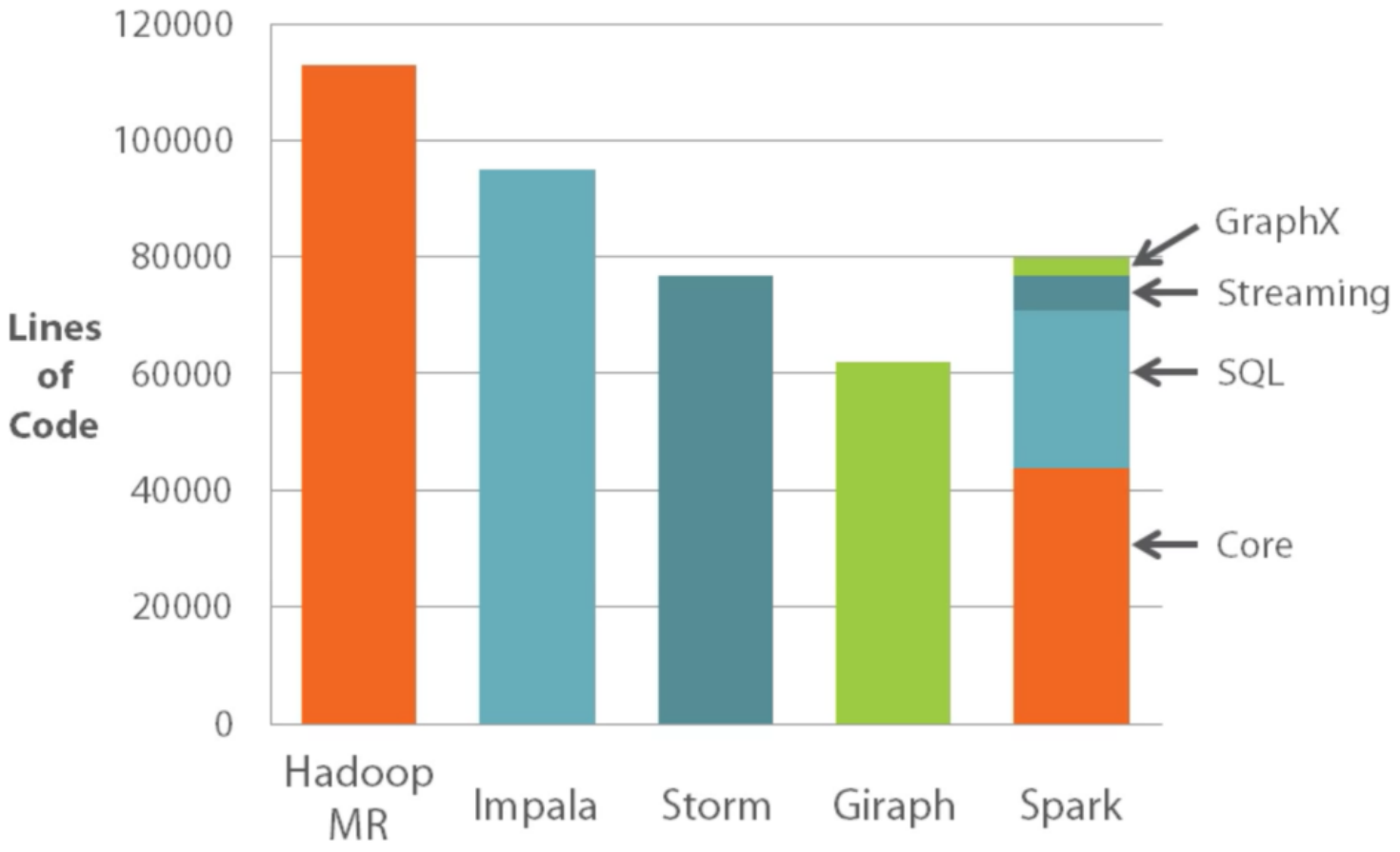
THE

# HISTORY OF SPARK



SPARK

# SHARED LIBRARIES

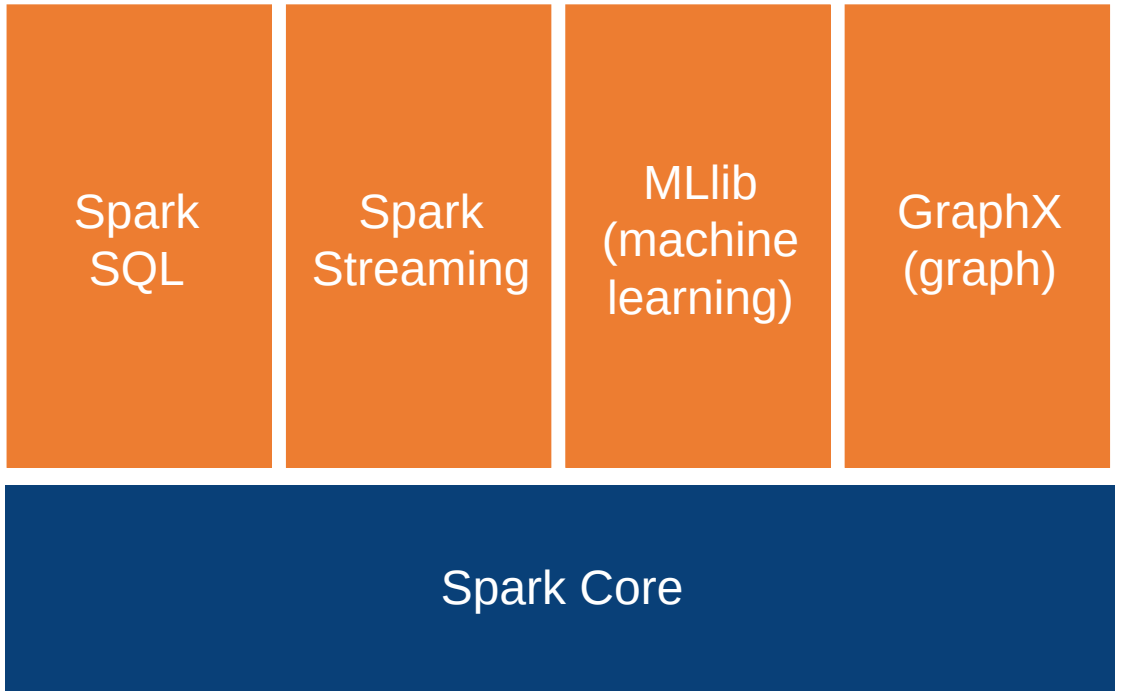


SPARK

# THE UNIFIED PLATFORM FOR BIG DATA

APIs for :

- Scala
- Java
- Python
- R



SPARK

# BENEFITS

## Performance

Using in-memory computing, Spark is considerably faster than Hadoop (100x in some tests).

Can be used for batch and real-time data processing.

## Developer Productivity

Easy-to-use APIs for processing large datasets.

Includes 100+ operators for transforming.

## Unified Engine

Integrated framework includes higher-level libraries for interactive SQL queries, processing streaming data, machine learning and graph processing.

A single application can combine all types of processing.

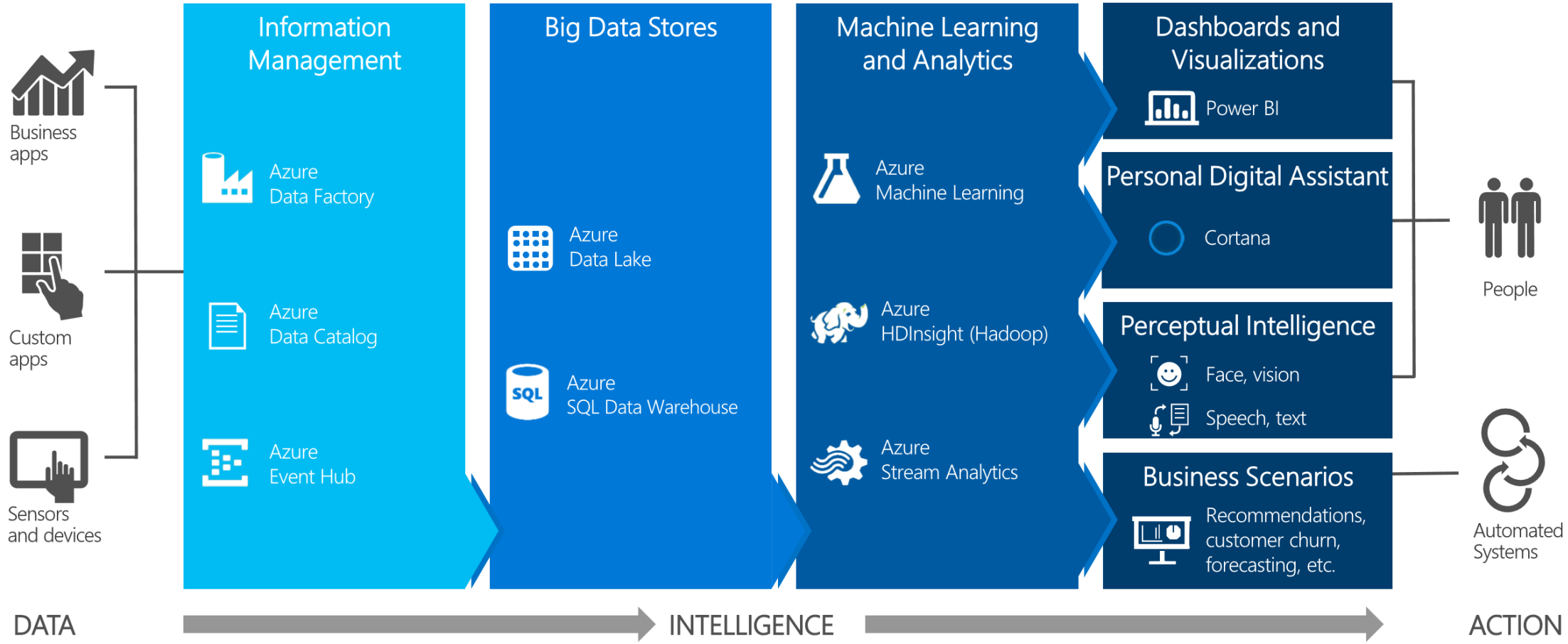
## Ecosystem

Spark has built-in support for many data sources such as HDFS, RDBMS, S3, Apache Hive, Cassandra and MongoDB.

Runs on top of the Apache YARN resource manager.

ANALYTICS

# CORTANA




Microsoft Partner of the Year
2015 Winner
Big Data and Analytics

NEUDESIC

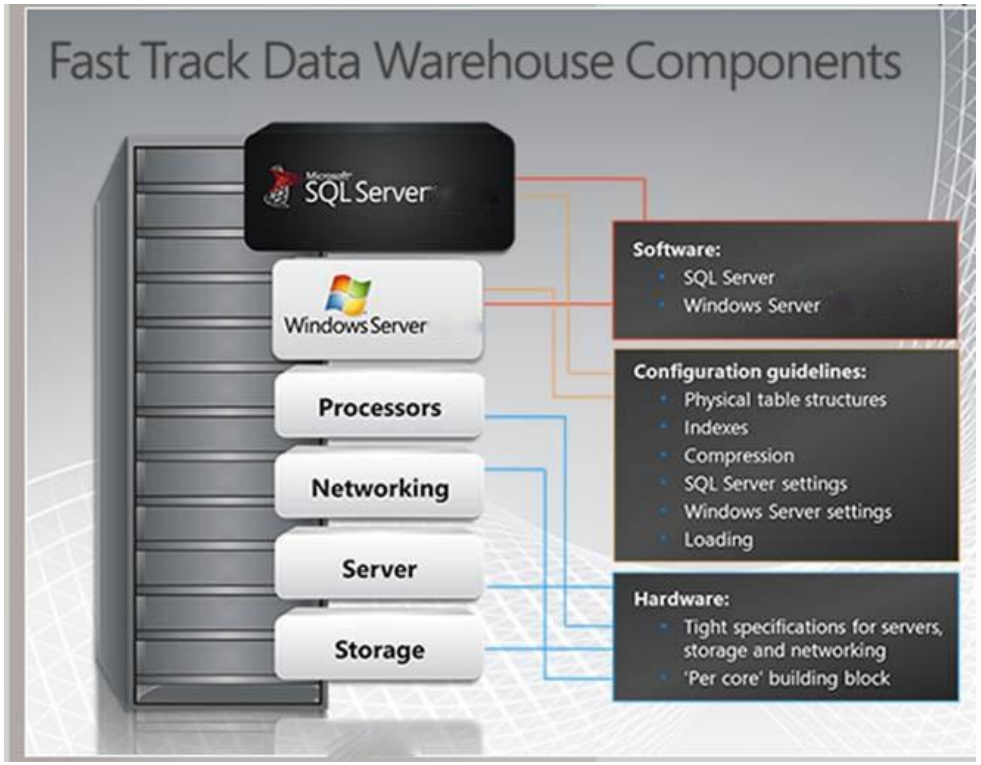
SQL Server

# BIG DATA OPTIMIZATIONS

Box Software

SQL Server

SQL Server Fast Track

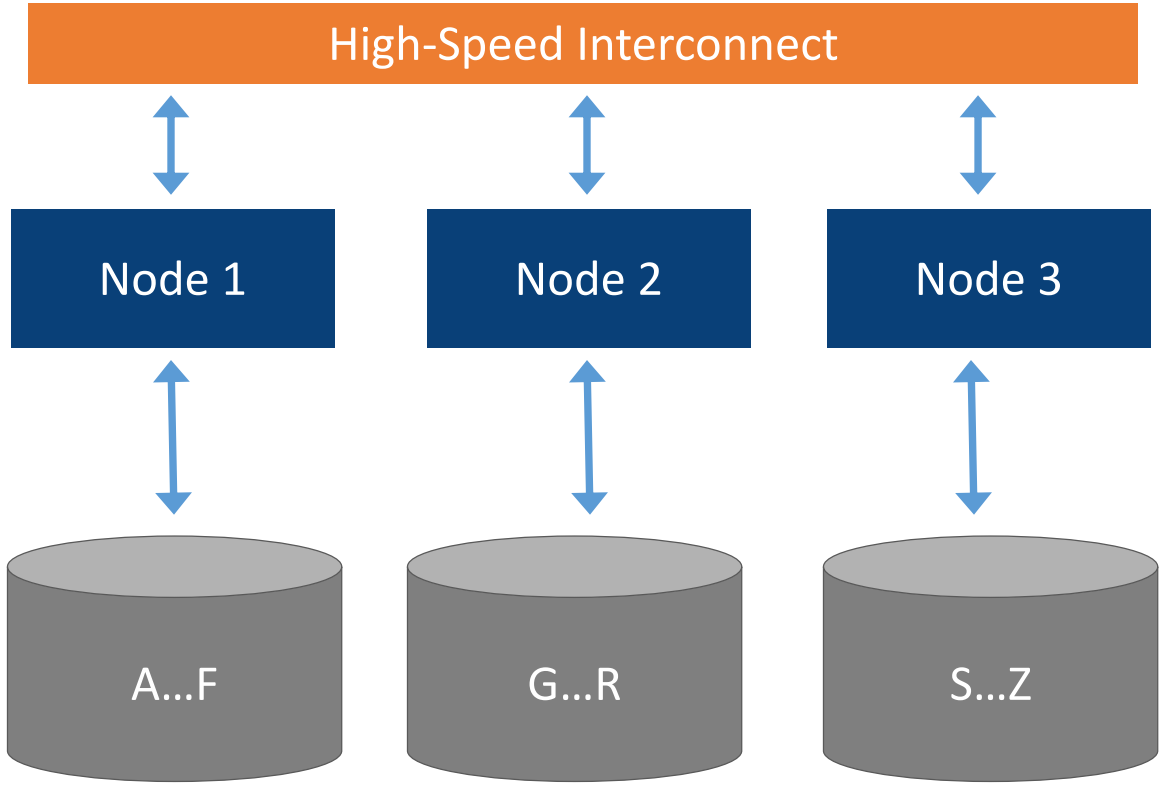



Microsoft Partner of the Year
2015 Winner
Big Data and Analytics

NEUDESIC

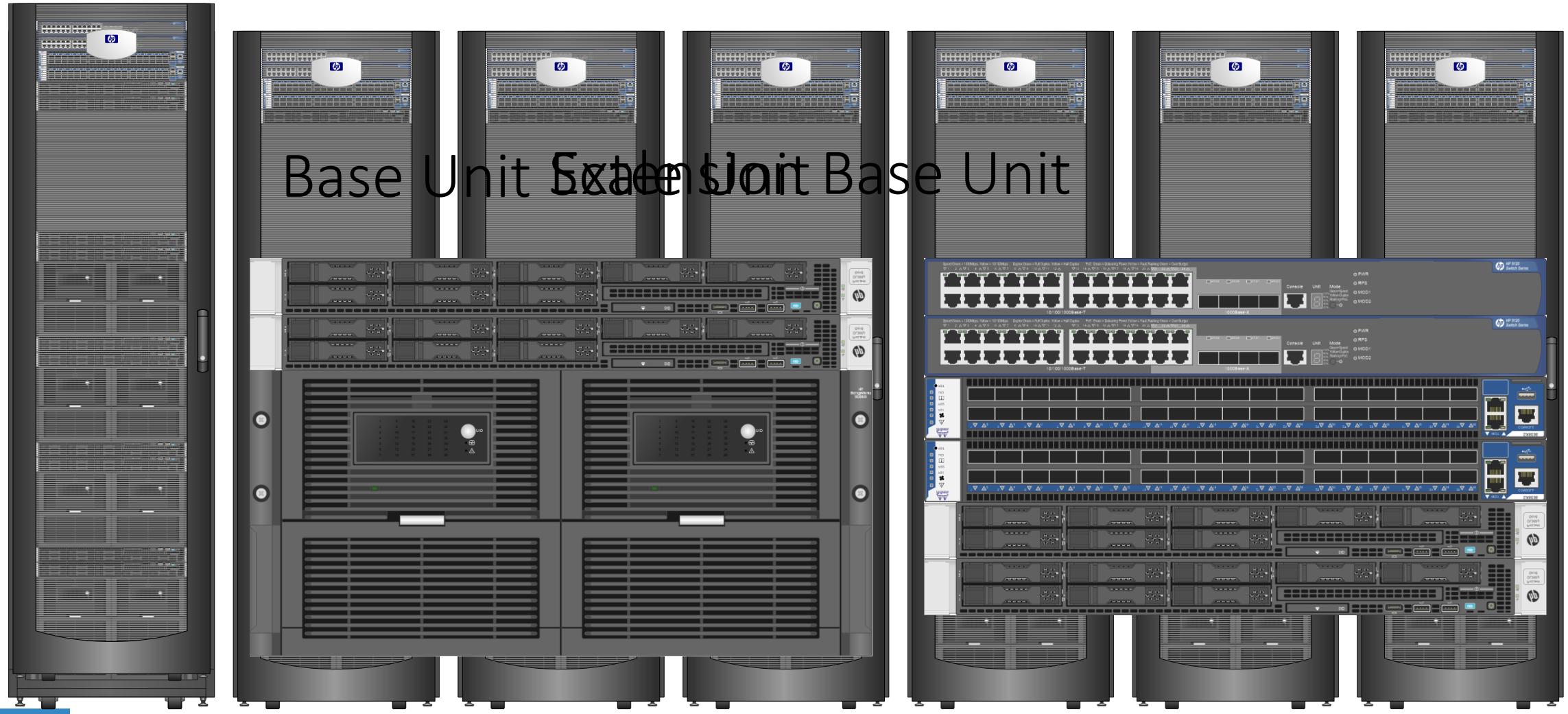
SQL Server

# APS



SQL Server

# APS GROWTH TOPOLOGY

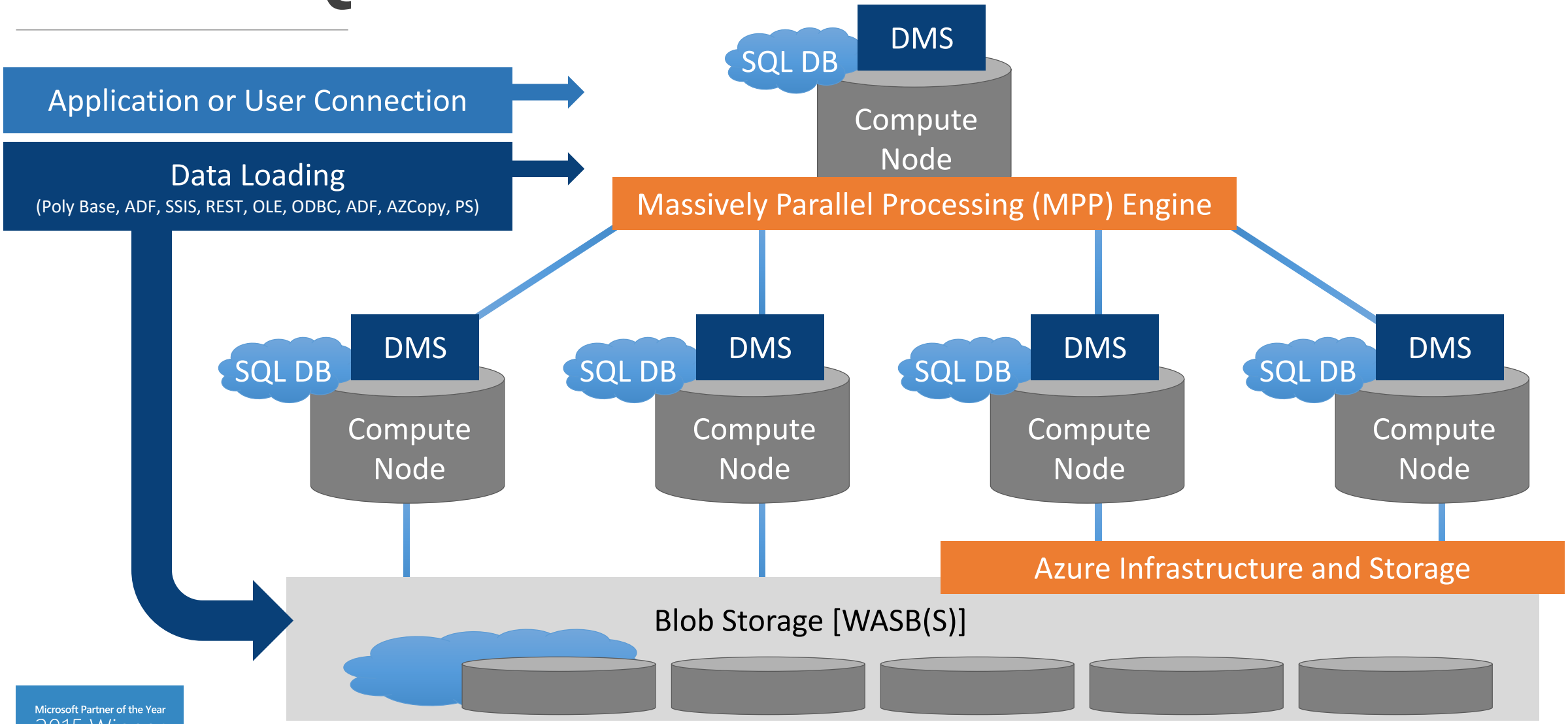



Microsoft Partner of the Year
2015 Winner
Big Data and Analytics

NEUDESIC

SQL Server

# Azure SQL DW




Microsoft Partner of the Year
2015 Winner
Big Data and Analytics

NEUDESIC

SQL Server

# DEPLOY OPTIONS & HYBRID SOLUTIONS

## Box Software

- SQL Server
- SQL Server Fast Track
- Hortonworks Data Platform

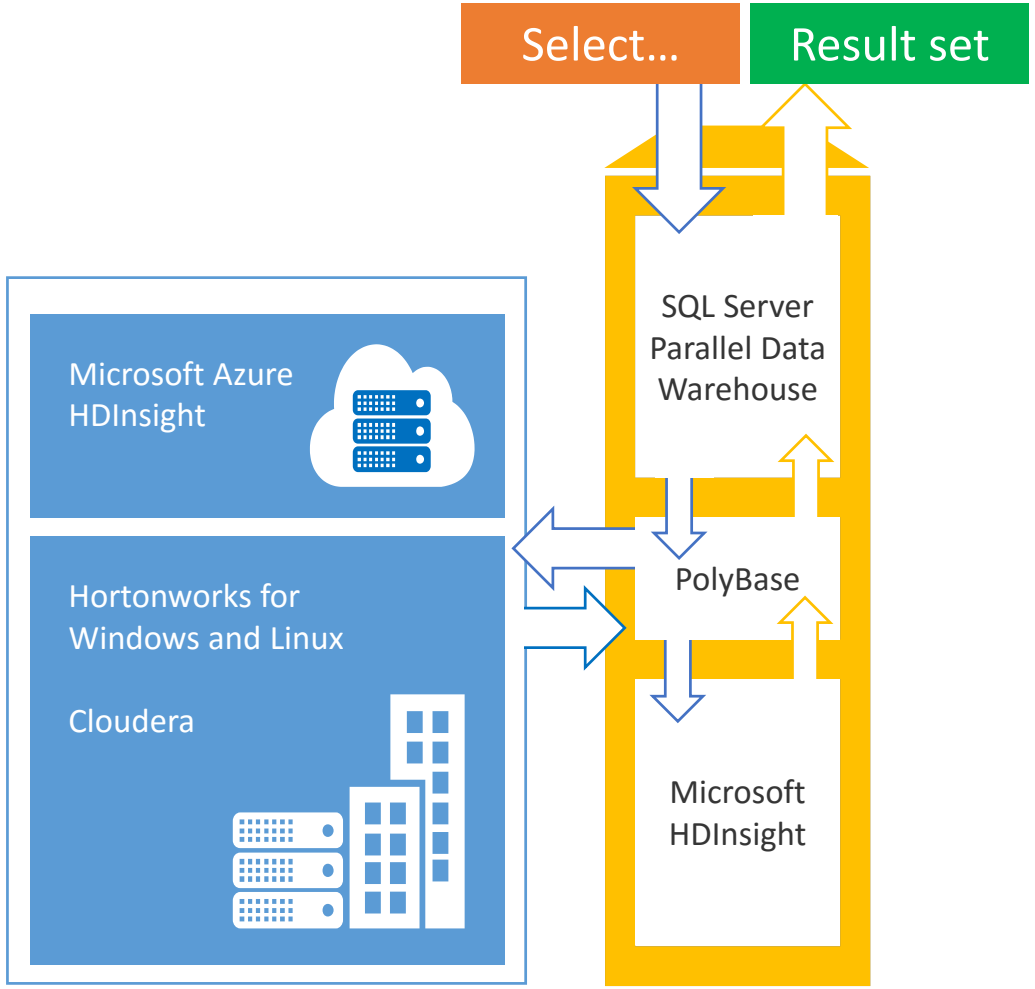
## Appliances

- Analytics Platform System

## Cloud

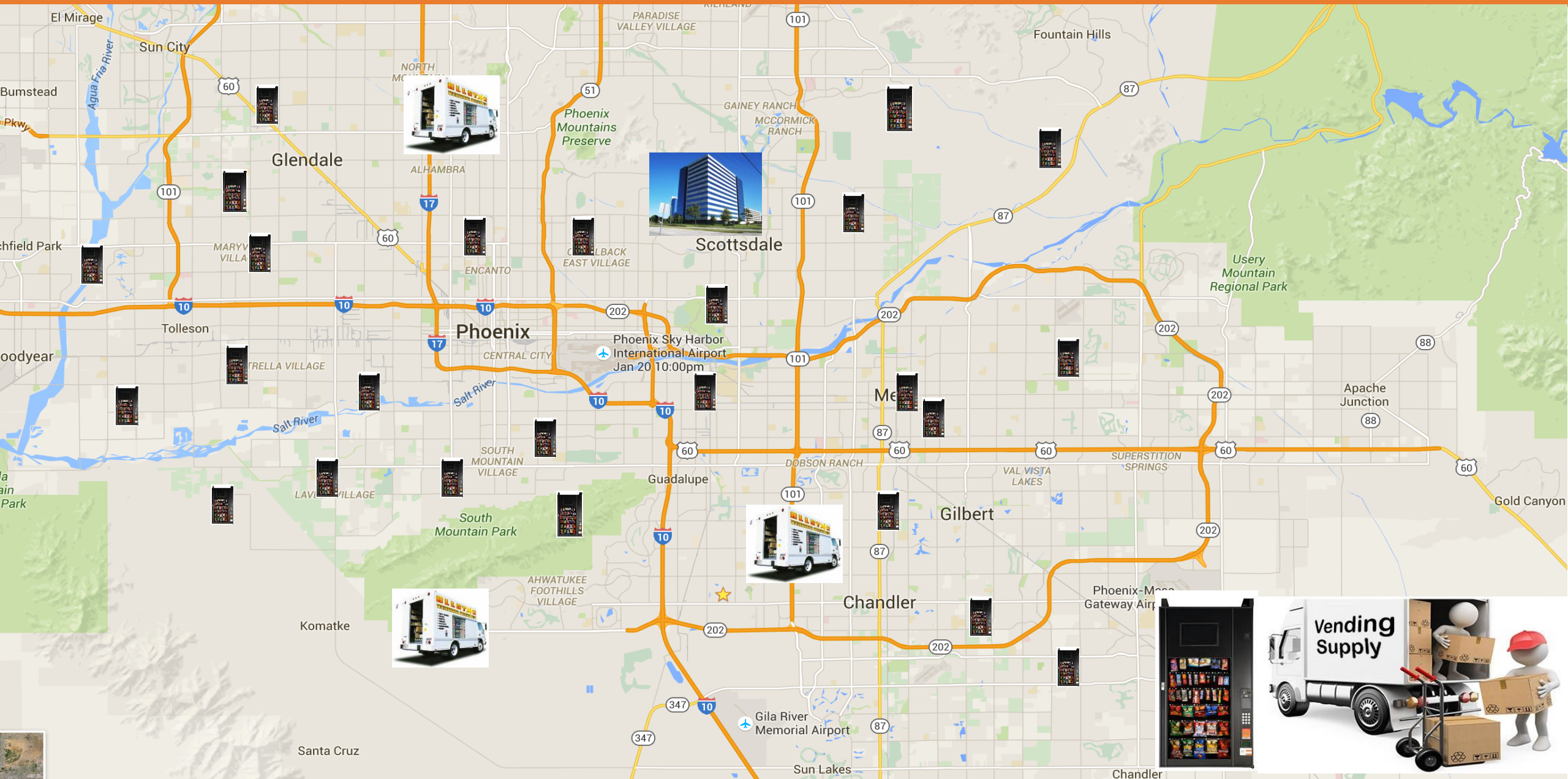
- Azure SQL DW
- HDInsight for Microsoft Azure

# CONNECTING ISLANDS OF DATA WITH POLYBASE



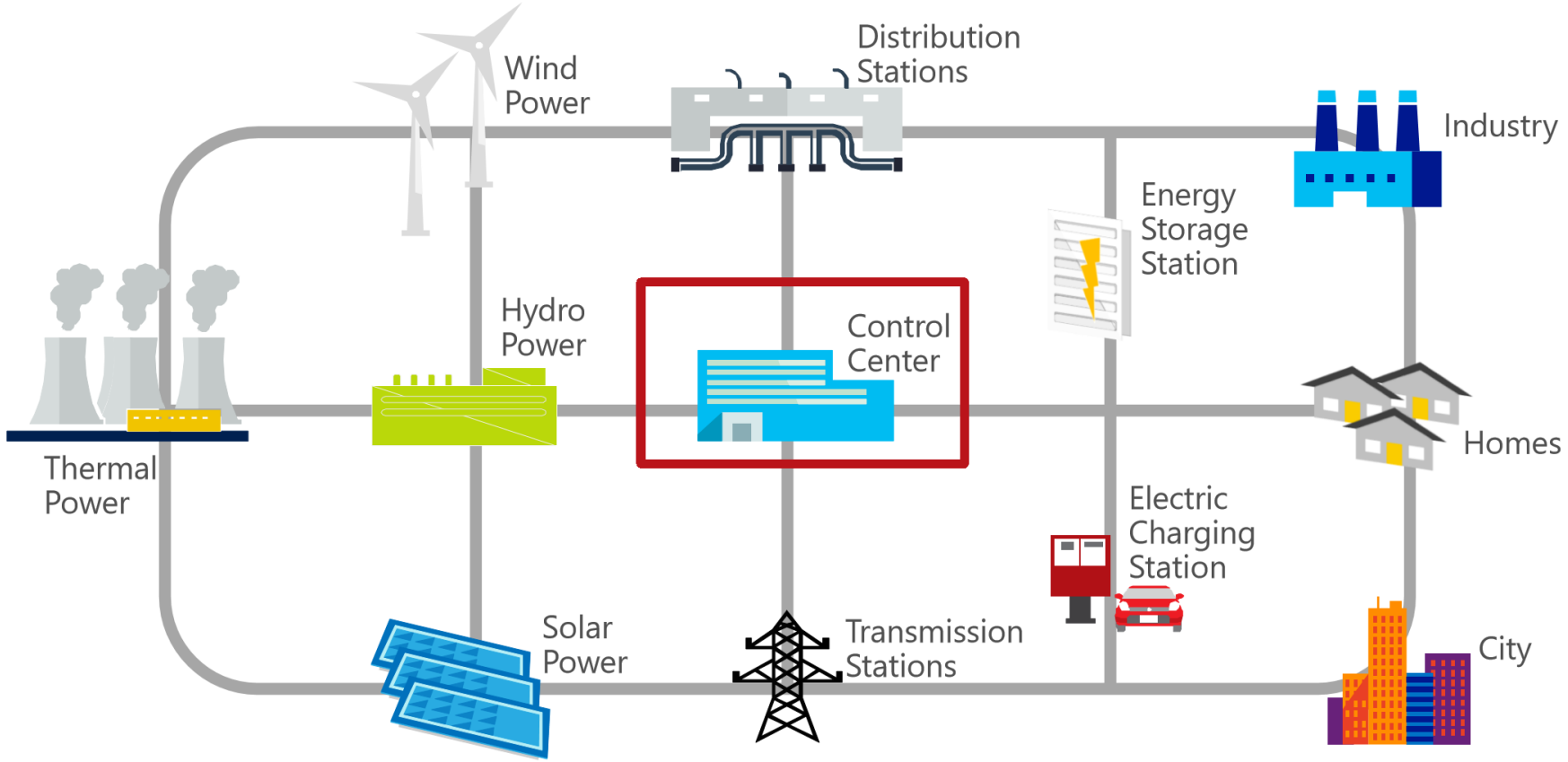

- Provides a single T-SQL query model for PDW and Hadoop with rich features of T-SQL, including joins without ETL
- Uses the power of MPP to enhance query execution performance
- Supports Windows Azure HDInsight to enable new hybrid cloud scenarios
- Provides the ability to query non-Microsoft Hadoop distributions, such as Hortonworks and Cloudera

# USE CASE: SUPPLY CHAIN MANAGEMENT



USE CASE:

# SMART GRID MANAGEMENT



USE CASES

# SMART GRID

PREDICTIVE MAINTENANCE

DEMAND FORECASTING

GRID OPTIMIZATION

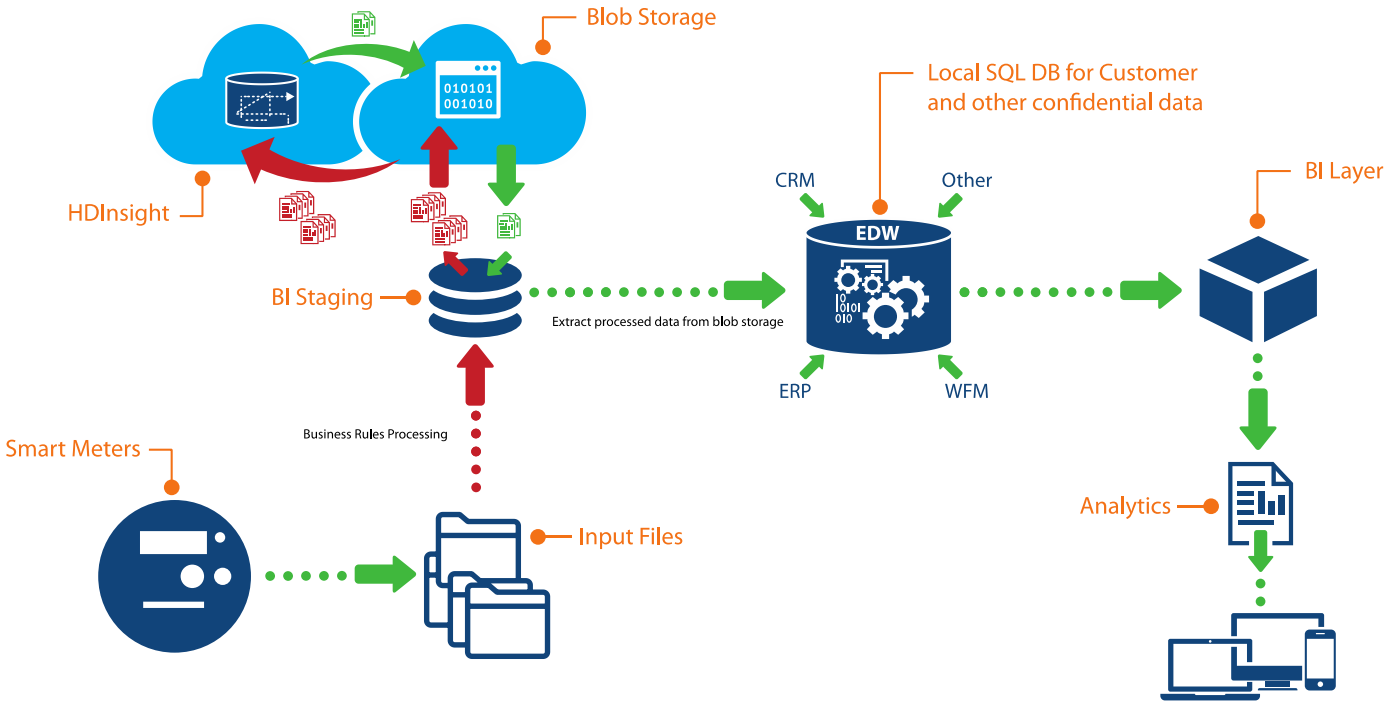
THEFT PERVENTION

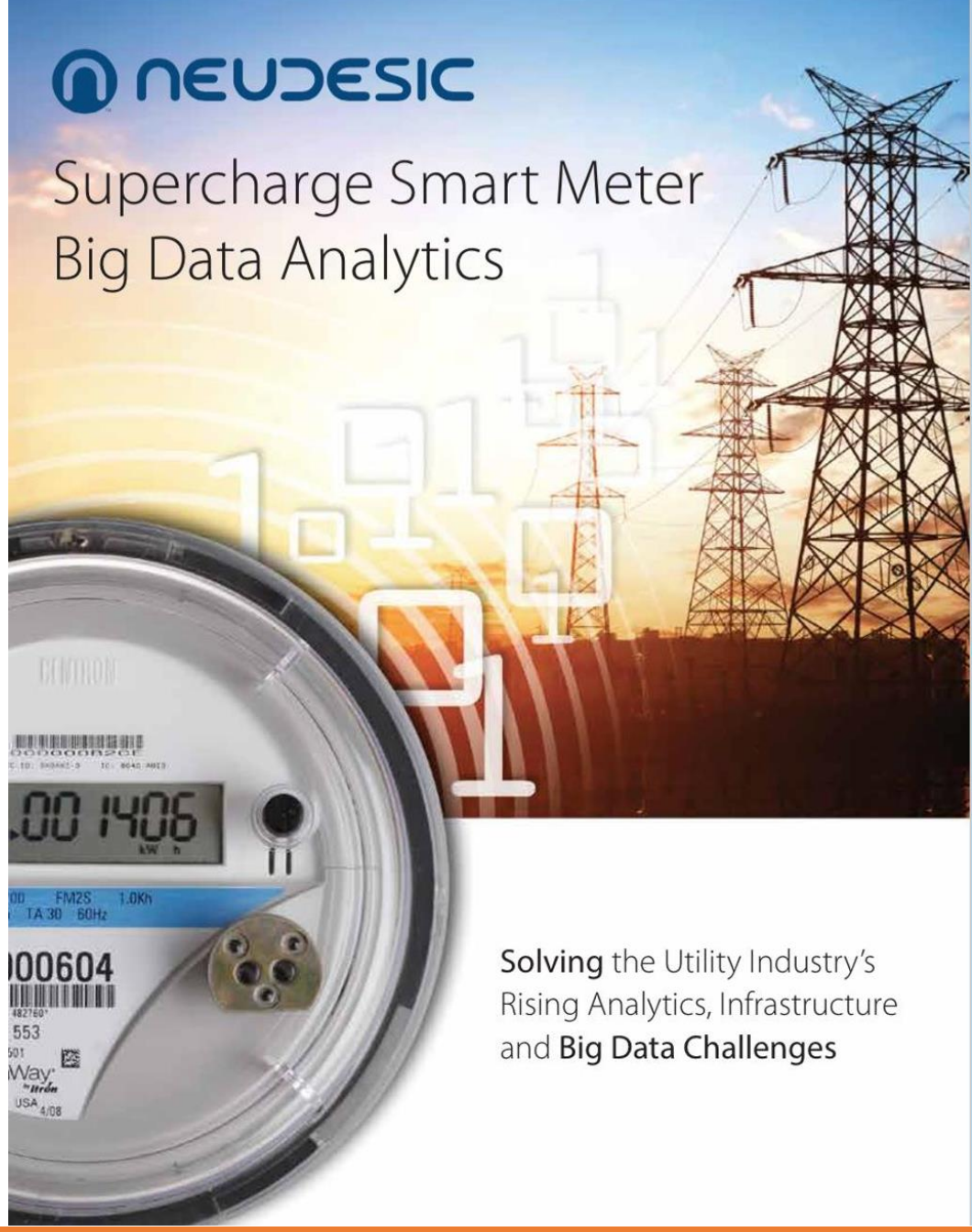
DEMAND RESPONSE

CUSTOMER PROFILING

**Neudesic** partnered with one of the nation's largest utility companies that recently deployed Smar Utility Meters for power customers, nearly a million meters sending usage data every 15 minutes.

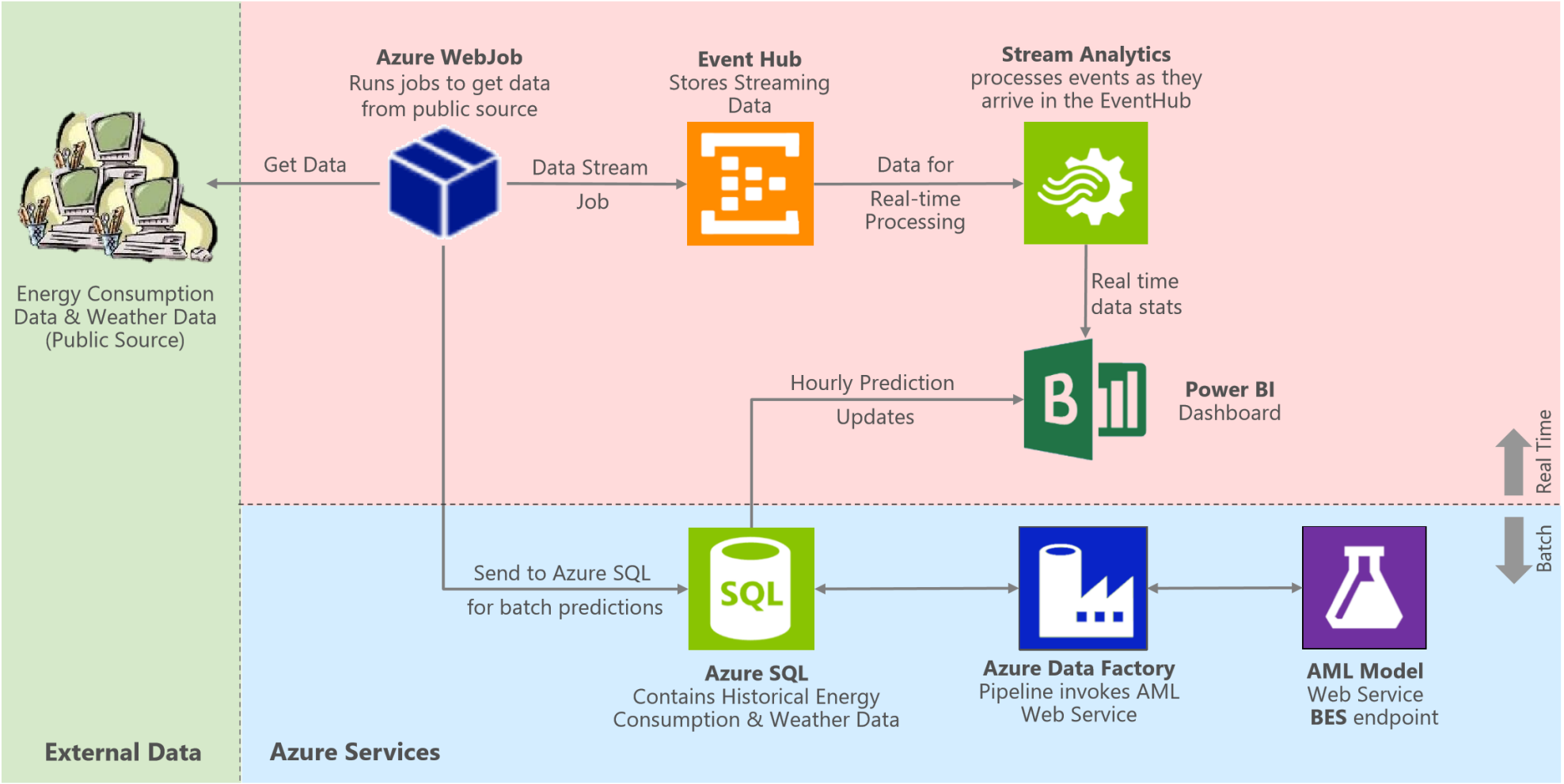
**The result:** an Azure hybrid big data processing solution that enabled the customer to perform gap analytics: a process for identifying gaps that exist in the power usage readings, over 7x faster than their previous solution! Billions of Smart Meter reads get processed to identify the nature and duration of the gaps to mitigate revenue losses.



USE CASES

# SMART GRID

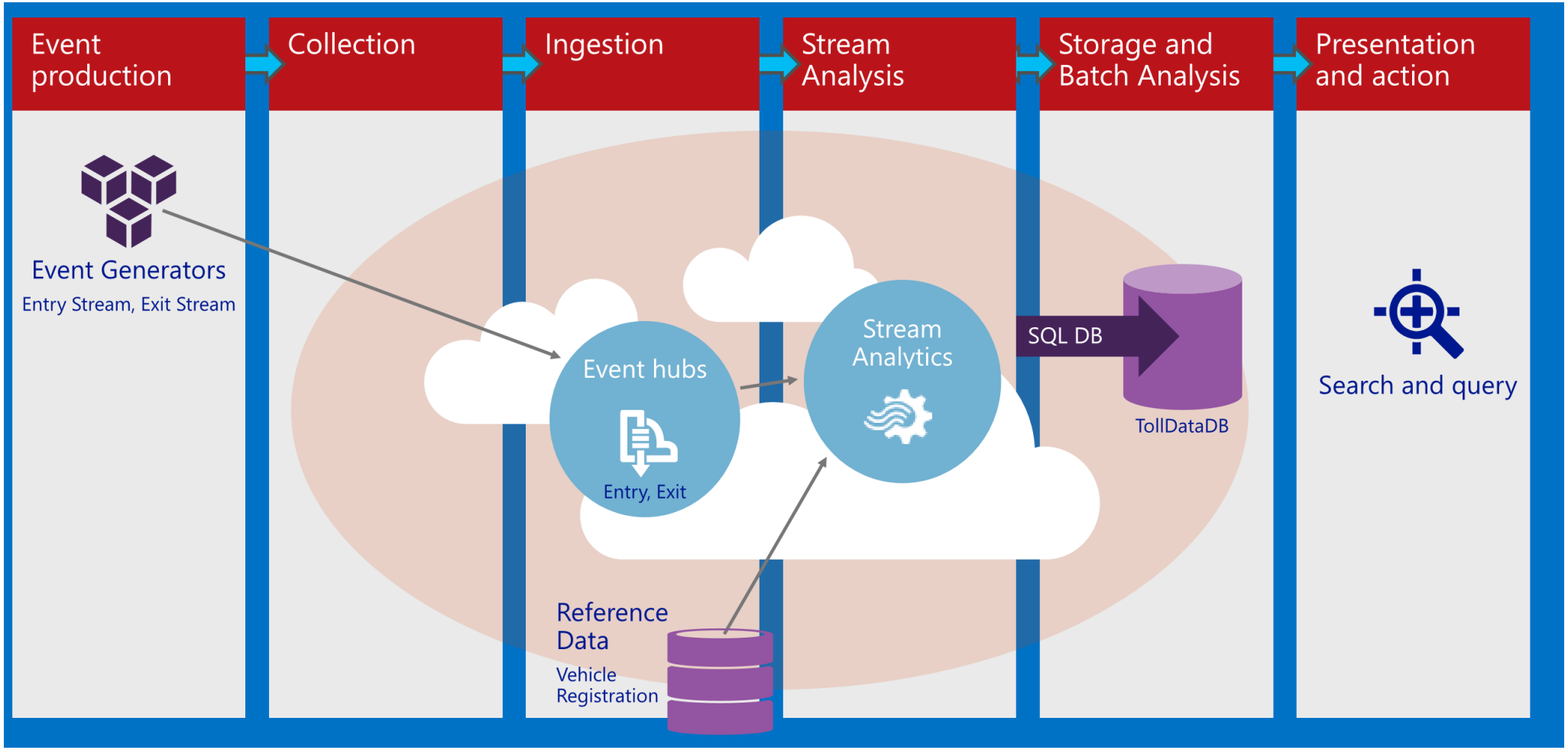



USE CASE:

# REAL TIME TRAFFIC ANALYSIS

REAL TIME

# TRAFFIC ANALYSIS



USE CASES

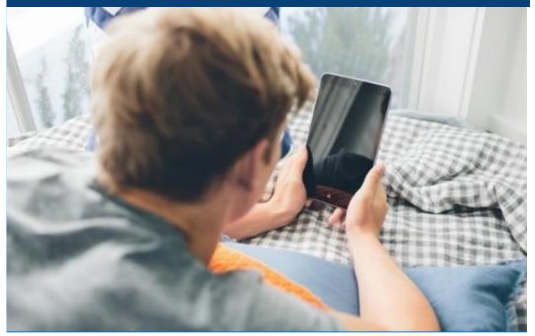
# STREAM ANALYTICS

**Real-time fraud detection**

**Connected cars**

**Click-stream analysis**

**Real-time financial portfolio alerts**



**Smart grid, energy management**


**CRM alerting sales to customer case**

**Data and identity protection services**

**Real-time sales tracking**

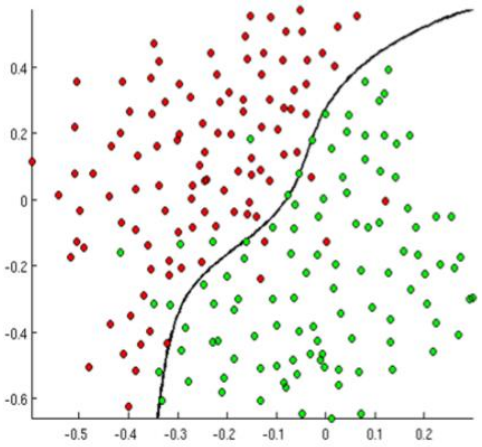
Microsoft Partner of the Year
2015 Winner
Big Data and Analytics

NEUDESIC

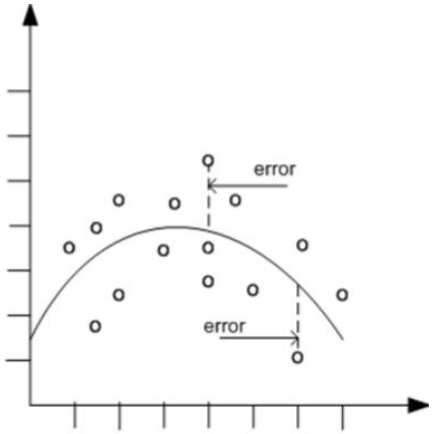
ML PROBLEMS

# SOLVED BY AZURE ML

Classification

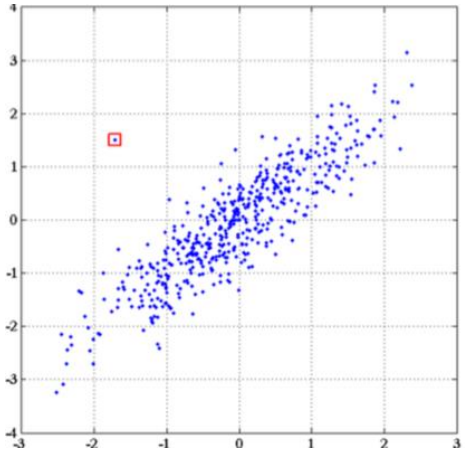



Regression

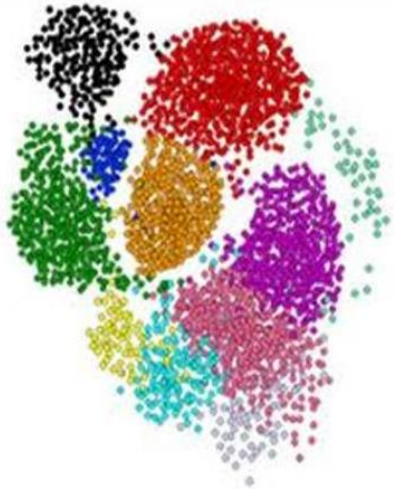



Recommenders


Anomaly Detection




Clustering



INDUSTRY USE CASES

# MACHINE LEARNING

## Financial Services

- New account risk screens
- Fraud prevention
- Trading risk
- Maximize deposit spread
- Insurance underwriting
- Accelerate loan processing

## Retail

- 360° view of the customer
- Analyze brand sentiment
- Localized, personalized promotions
- Website optimization
- Optimal store layout

## Telecom

- Call detail records (CDRs)
- Infrastructure investment
- Next product to buy (NPTB)
- Real-time bandwidth allocation
- New product development

## Manufacturing

- Supplier consolidation
- Supply chain and logistics
- Assembly line quality assurance
- Proactive maintenance
- Crowd source quality assurance

## Healthcare

- Genomic data for medical trials
- Monitor patient vitals
- Reduce re-admittance rates
- Store medical research data
- Recruit cohorts for pharmaceutical trials

## Utilities & Energy

- Smart meter stream analysis
- Slow oil well decline curves
- Optimize lease bidding
- Compliance reporting
- Proactive equipment repair
- Seismic mage processing

## Public Sector

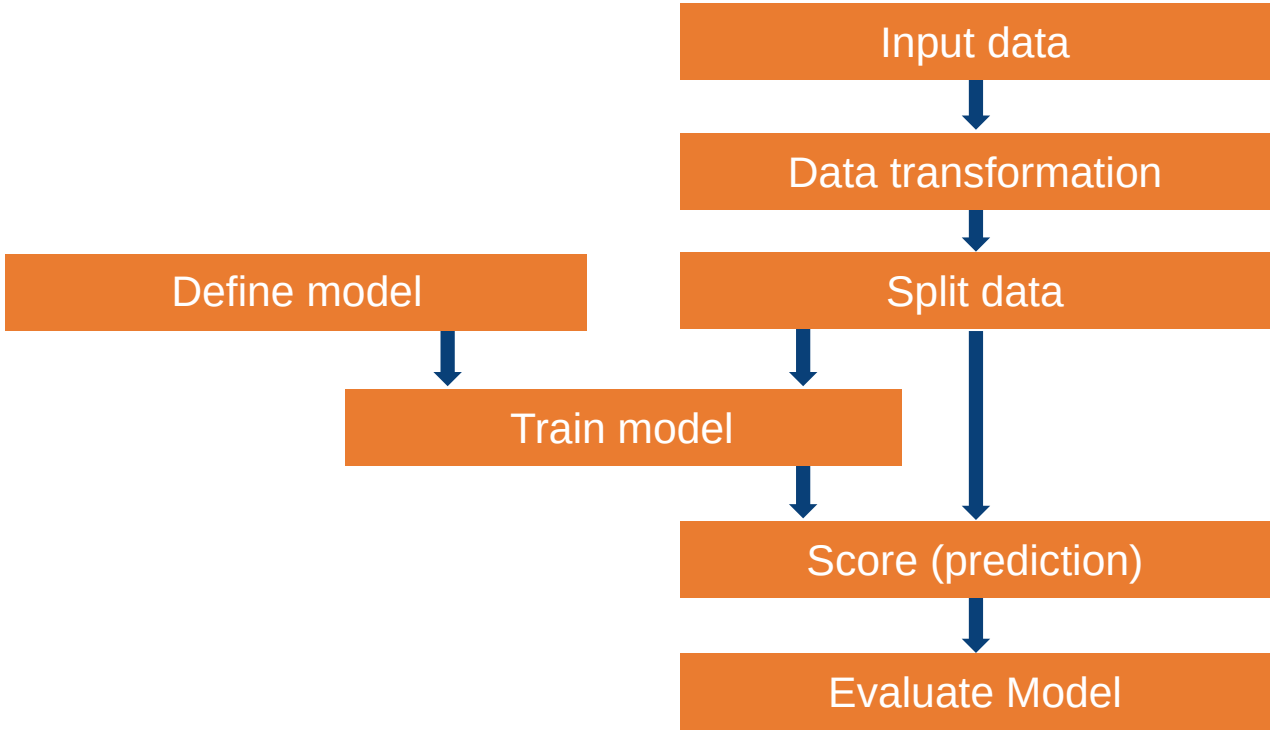
- Analyze public sentiment
- Protect critical networks
- Prevent fraud and waste
- Crowd source reporting for repairs to infrastructure
- Fulfill open records requests

## Goods and Manufacturing

- Consumer Goods & Identify hidden revenue opportunities
- See and predict changes in supply or demand Market price volatility and production planning Promotional demand Suggested product engines

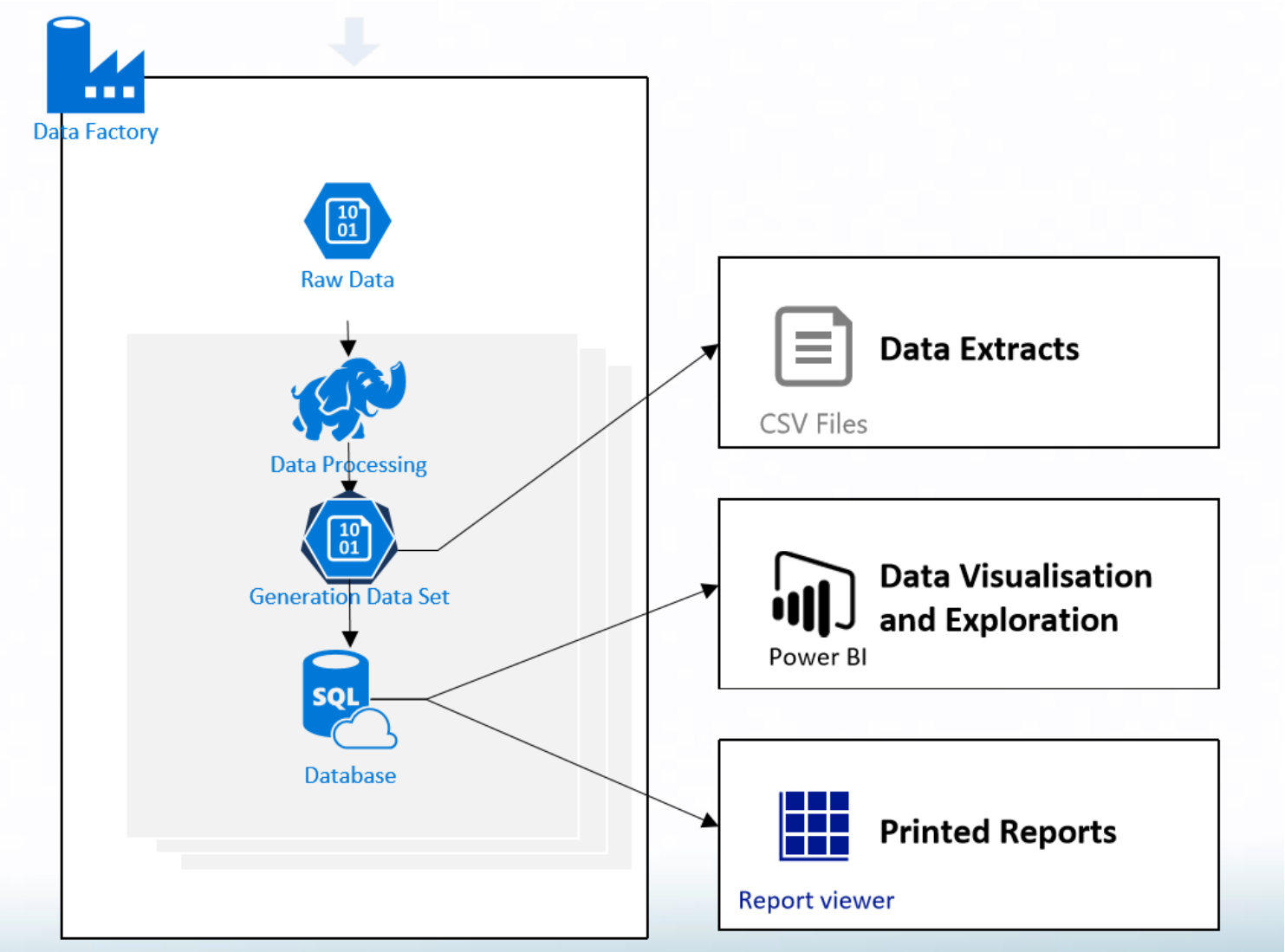
THE

# MACHINE LEARNING WORKFLOW



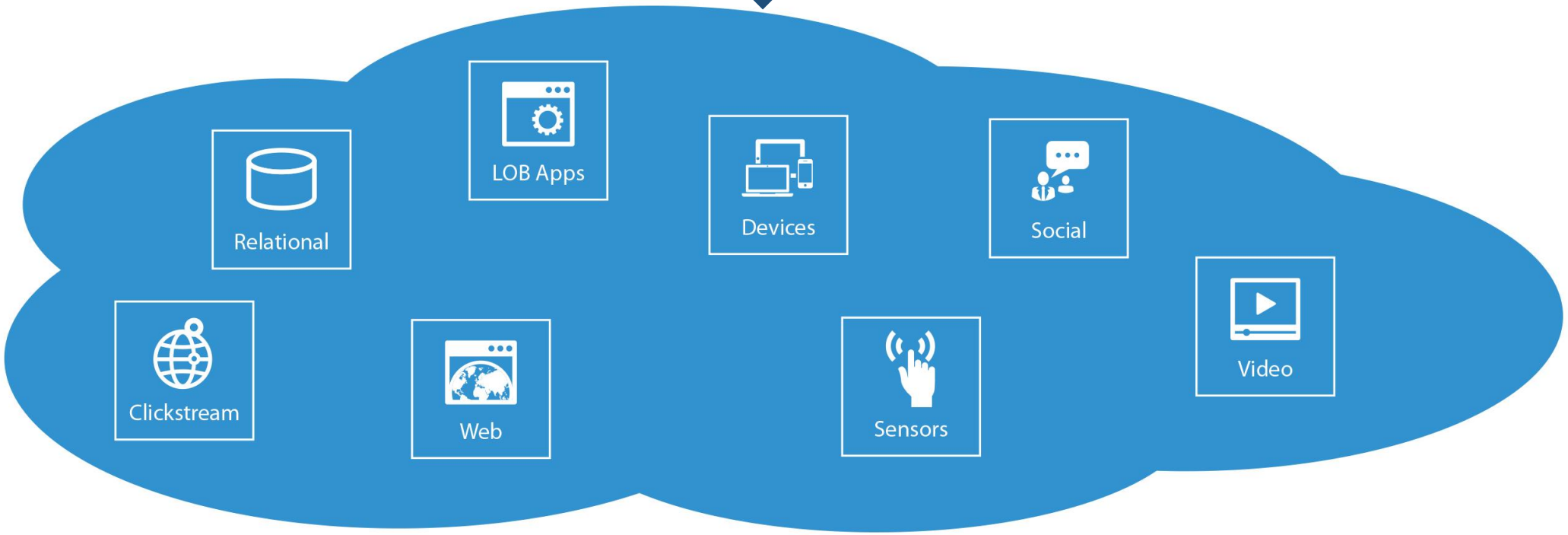
AZURE

# DATA FACTORY



HD

# INSIGHT


Microsoft Partner of the Year
2015 Winner
Big Data and Analytics

NEUDESIC

BIG DATA &

# Advanced Analytics Roadshow

*Questions?*

Orion Gebremedhin
Orion.Gebremedhin@Neudesic.com
Twitter: @oriongm

Marc Lobree
Marc.Lobree@Neudesic.com

Microsoft Partner of the Year
2015 Winner
Big Data and Analytics

NEUDESIC