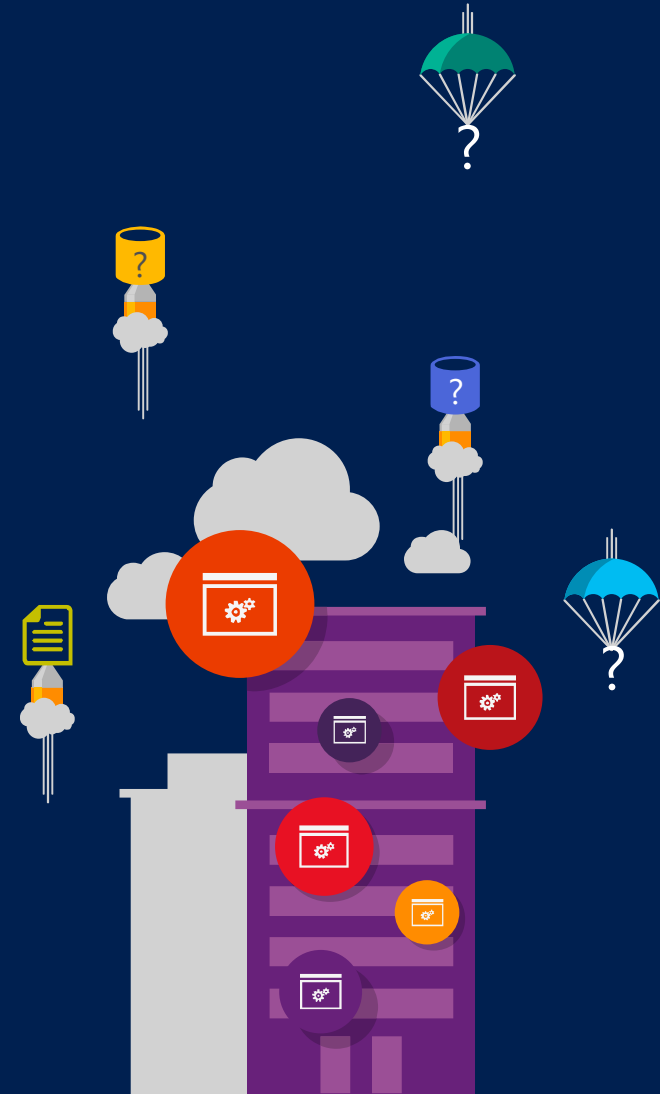


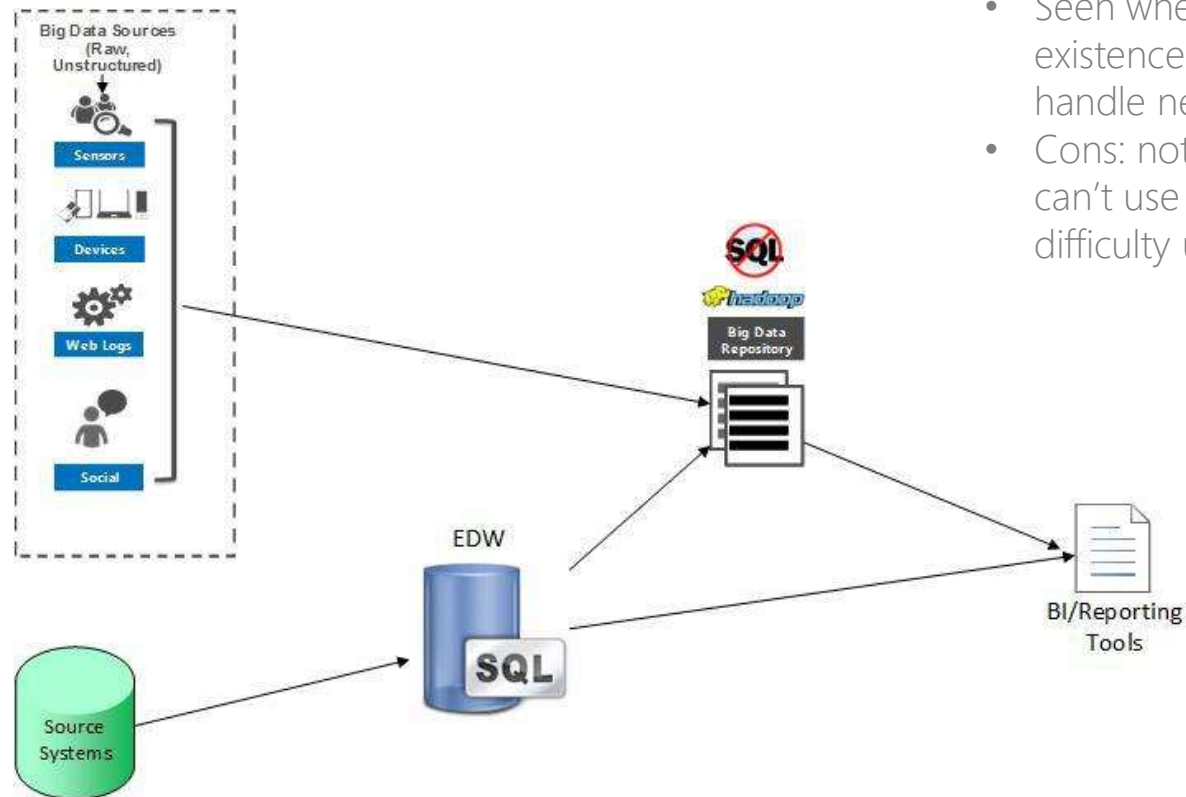
# Agenda

- Big Data Architectures
- Why data lakes?
- Top-down vs Bottom-up
- Data lake defined
- Hadoop as the data lake
- Modern Data Warehouse

# Big Data Architectures

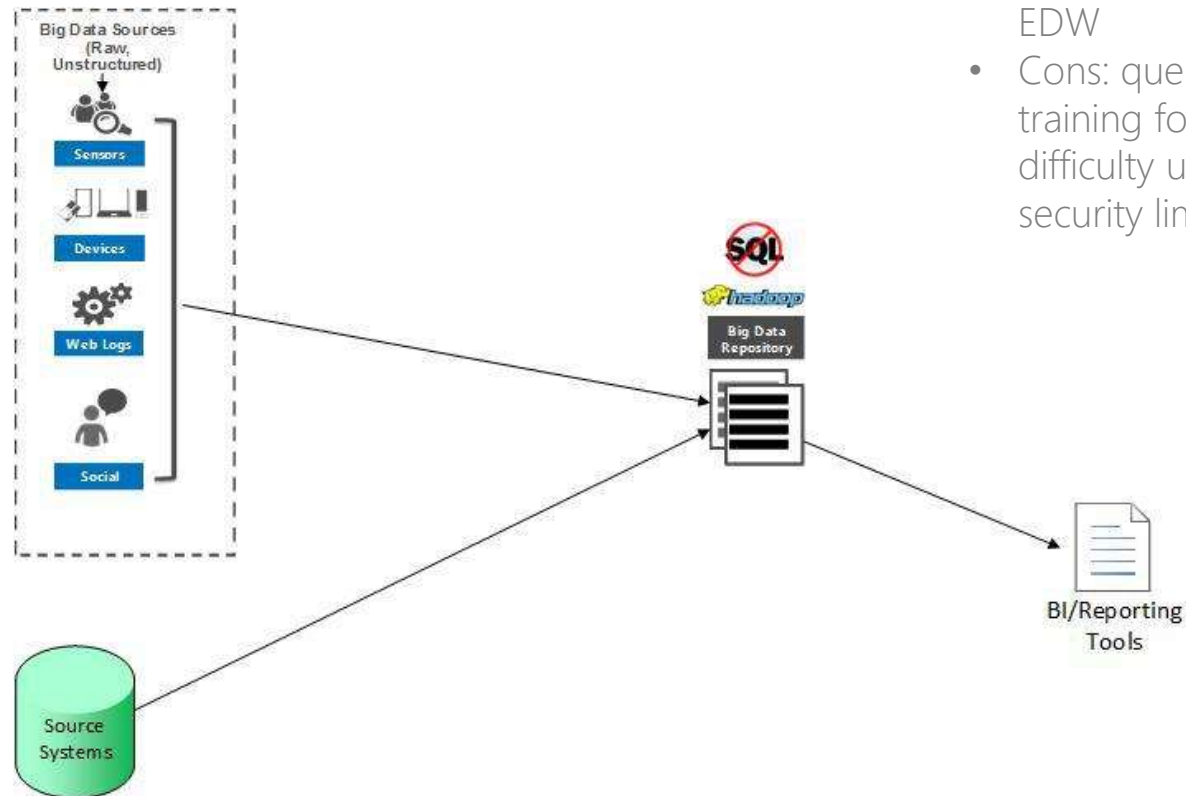


# Enterprise data warehouse augmentation



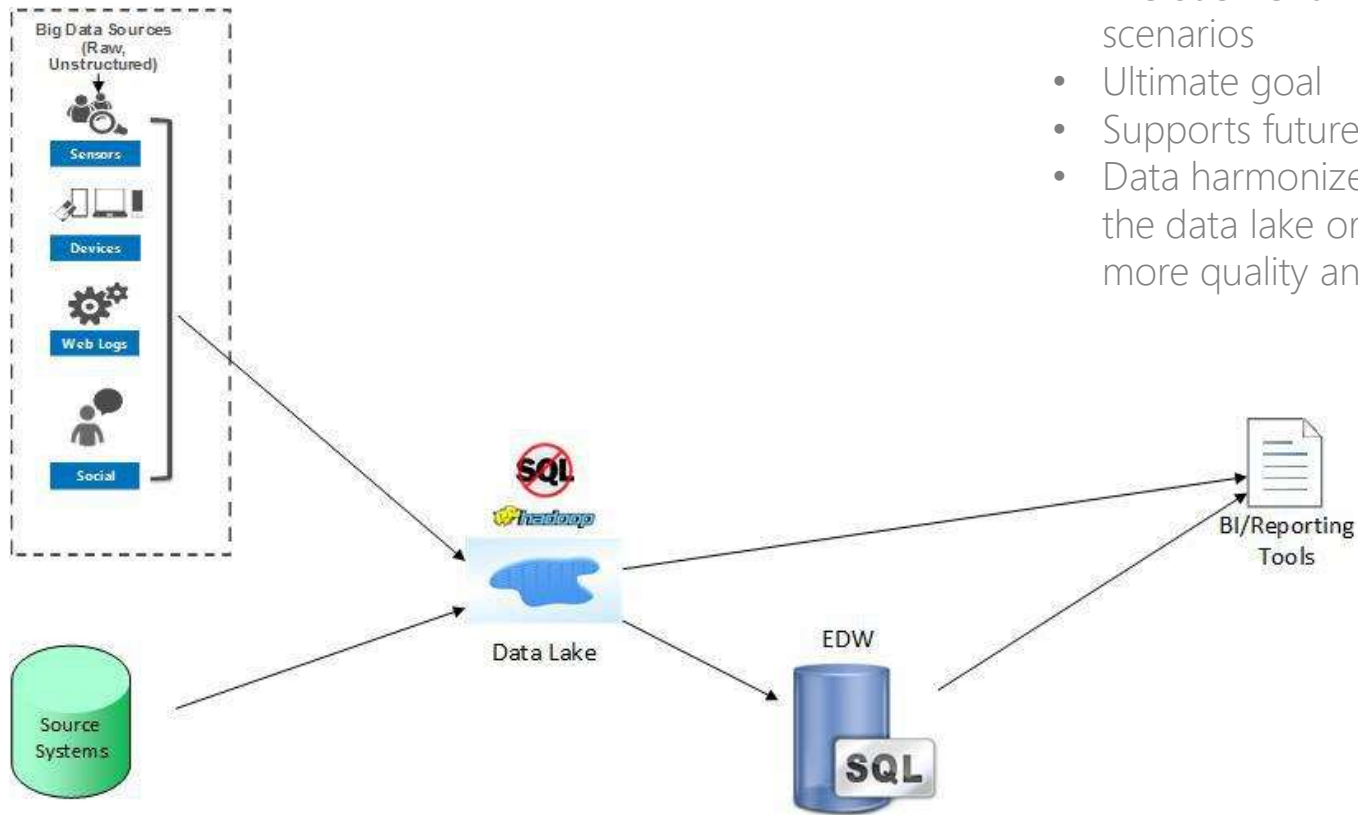
- Seen when EDW has been in existence a while and EDW can't handle new data
- Cons: not offloading EDW work, can't use existing tools, data hub difficulty understanding data

# All-in-one



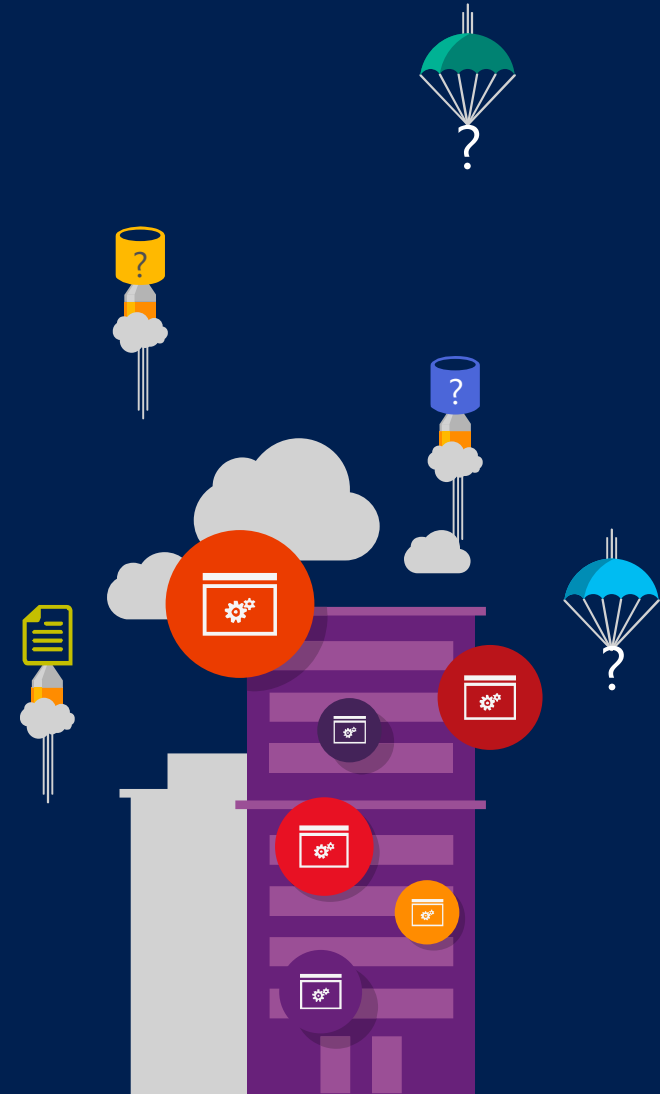
- Data hub is total solution, no EDW
- Cons: queries are slower, new training for reporting tools, difficulty understanding data, security limitations

# Modern Data Warehouse



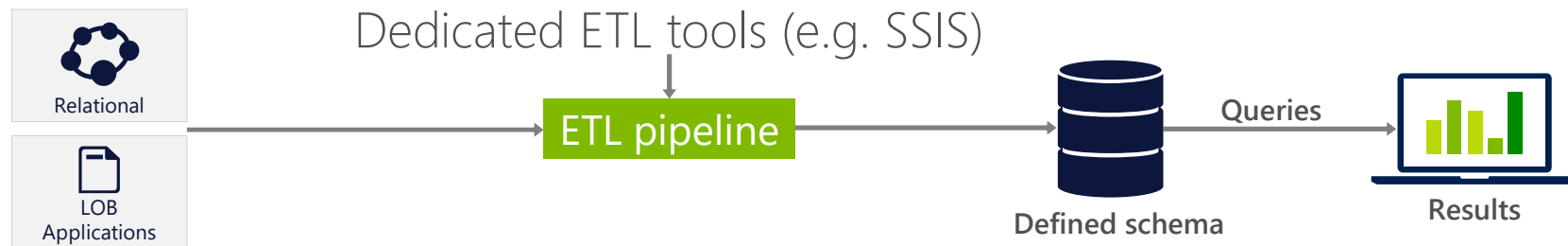
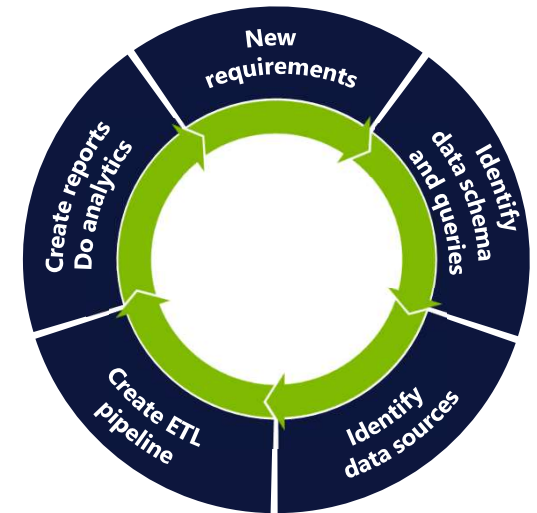
- Evolution of three previous scenarios
- Ultimate goal
- Supports future data needs
- Data harmonized and analyzed in the data lake or moved to EDW for more quality and performance

# Why data lakes?



# 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

# Need to collect any data

Harness the growing and changing nature of data

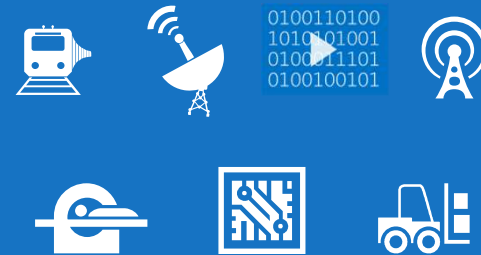
## Structured



## Unstructured



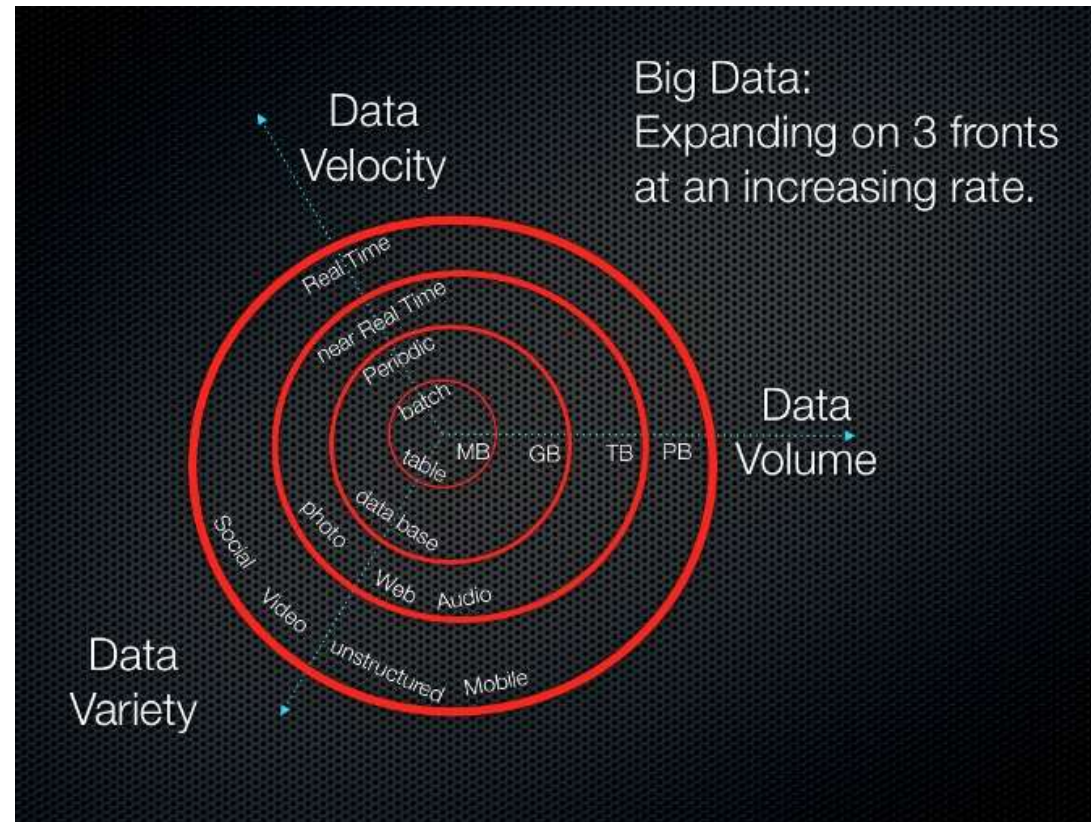
## Streaming



- ▶ Challenge is combining transactional data stored in relational databases with less structured data
- ▶ Big Data = All Data
- ▶ Get the right information to the right people at the right time in the right format



# The three V's












# New big data thinking: All data has value

- ⚡ All data has potential value
- ⚡ Data hoarding
- ⚡ 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



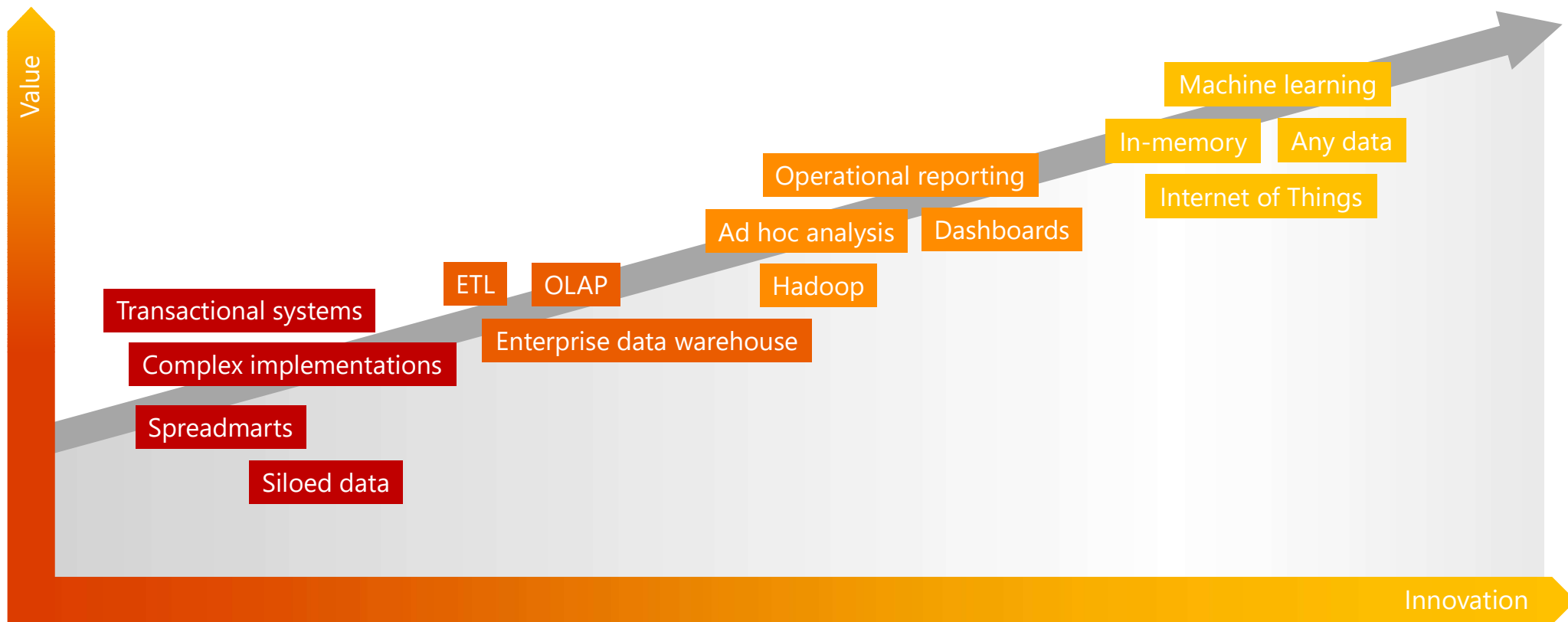
# Data Lake Store: Technical Requirements

	<b>Secure</b>	Must be highly secure to prevent unauthorized access (especially as all data is in one place).
	<b>Scalable</b>	Must be highly scalable. When storing all data indefinitely, data volumes can quickly add up
	<b>Reliable</b>	Must be highly available and reliable (no permanent loss of data).
	<b>Throughput</b>	Must have high throughput for massively parallel processing via frameworks such as Hadoop and Spark
	<b>Low latency</b>	Must have low latency for high-frequency operations.
	<b>Details</b>	Must be able to store data with all details; aggregation may lead to loss of details.
	<b>Native format</b>	Must permit data to be stored in its 'native format' to track lineage & for data provenance.
	<b>All sources</b>	Must be able ingest data from a variety of sources-LOB/ERP, Logs, Devices, Social NWs etc.
	<b>Multiple analytic frameworks</b>	Must support multiple analytic frameworks—Batch, Real-time, Streaming, ML etc. No one analytic framework can work for all data and all types of analysis.

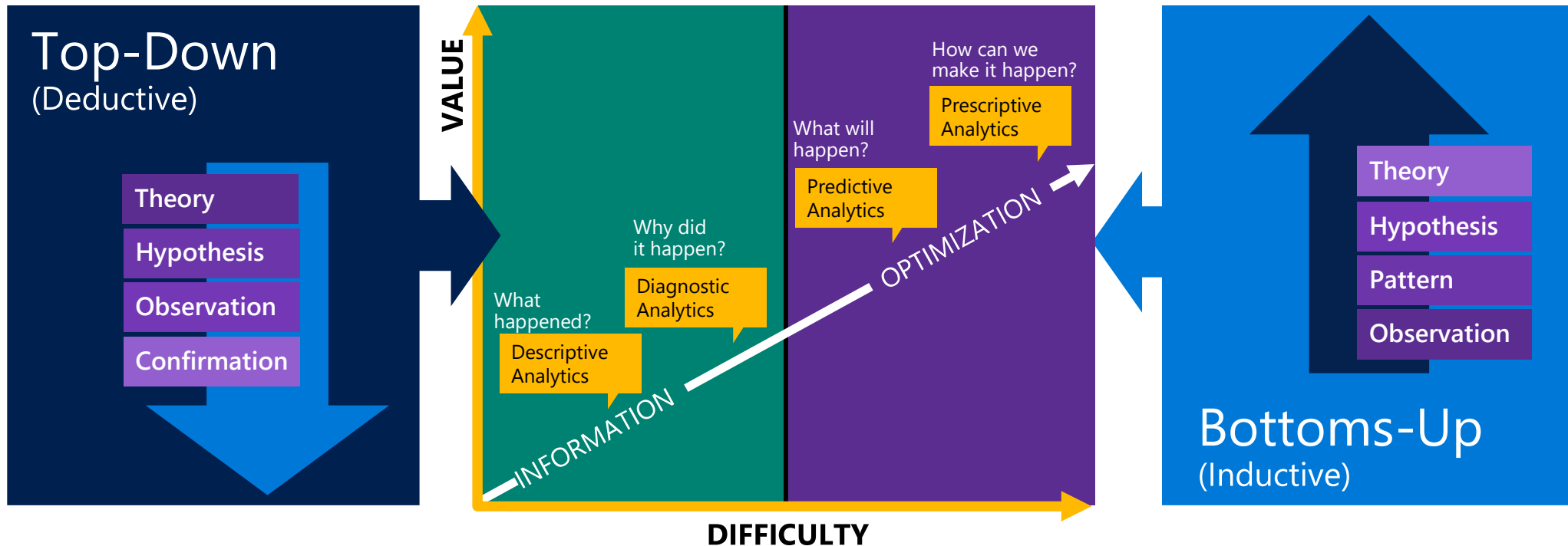
# Top-down vs Bottom-up



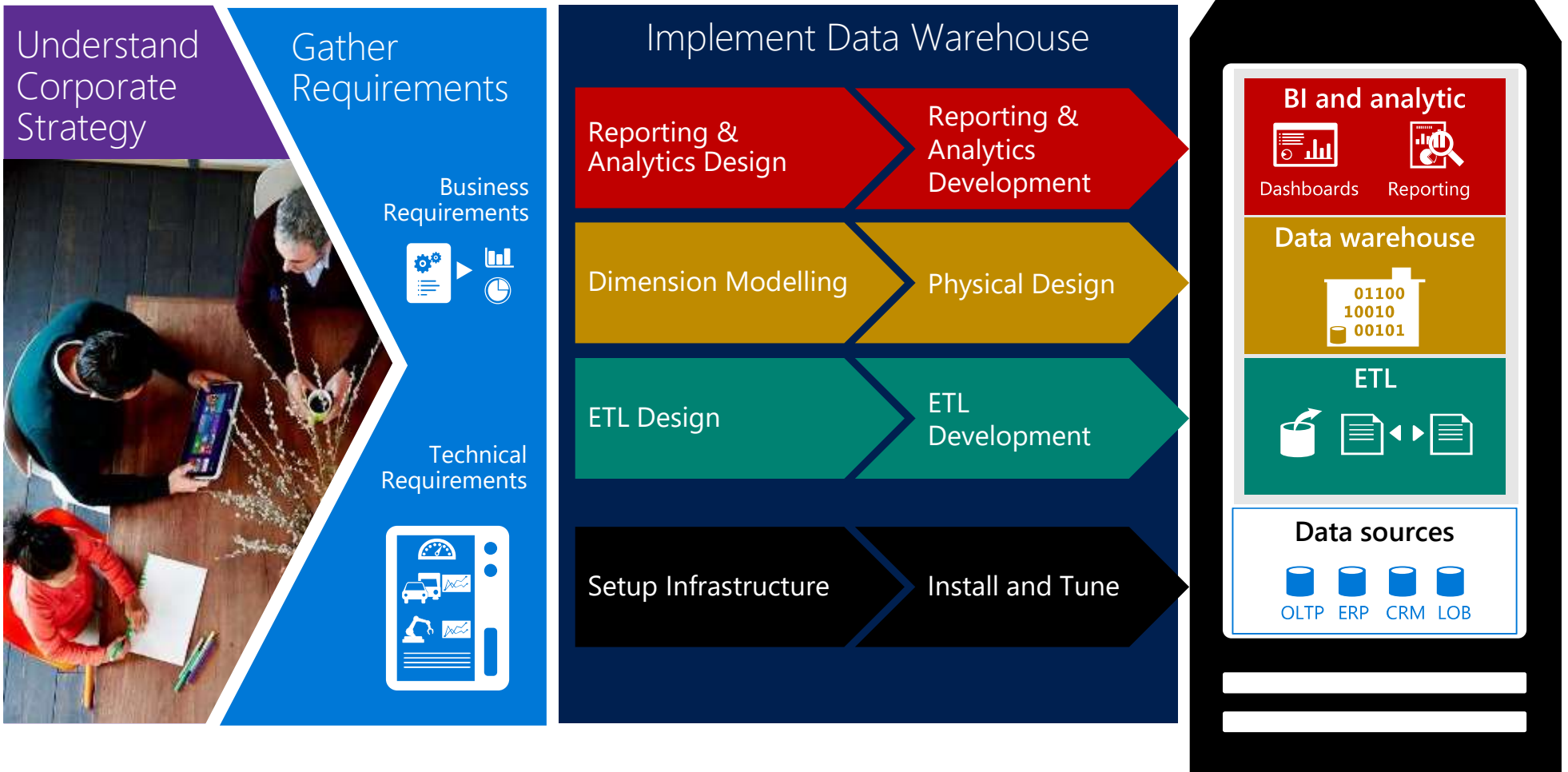
# Technology innovation accelerates value



# Two Approaches to Information Management for Analytics: Top-Down + Bottoms-Up



# Data Warehousing Uses A Top-Down Approach



# The “data lake” Uses A Bottoms-Up Approach





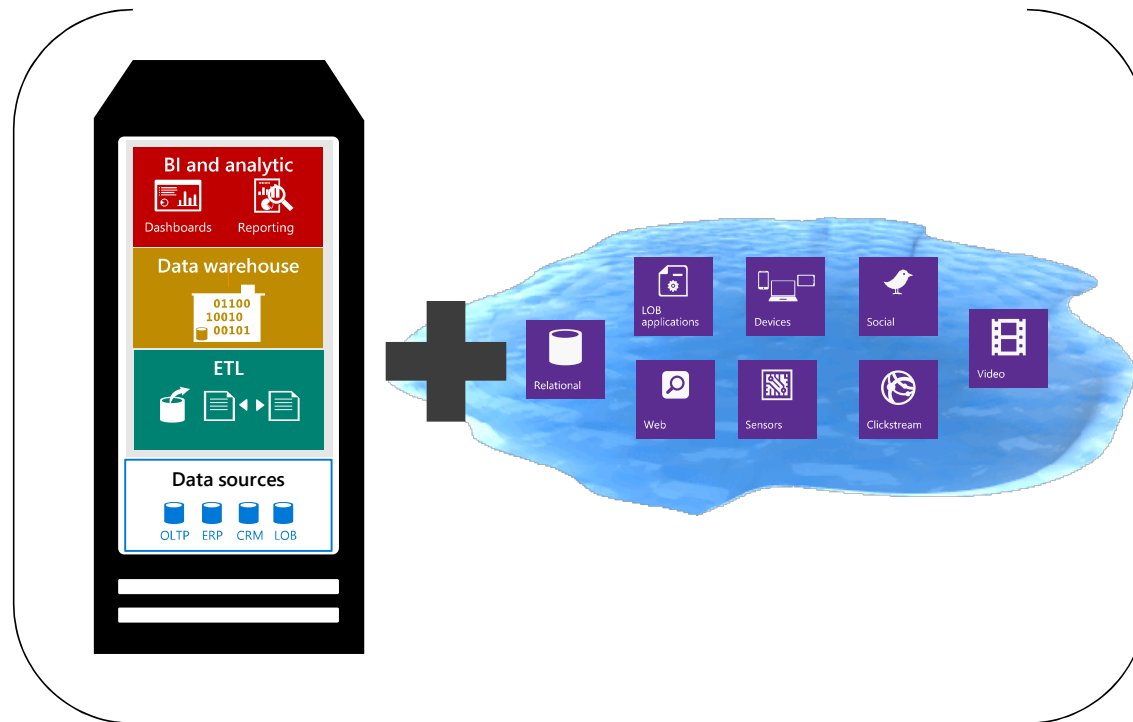
# Data Lake + Data Warehouse Better Together

What happened?

Descriptive  
Analytics

Why did it happen?

Diagnostic  
Analytics



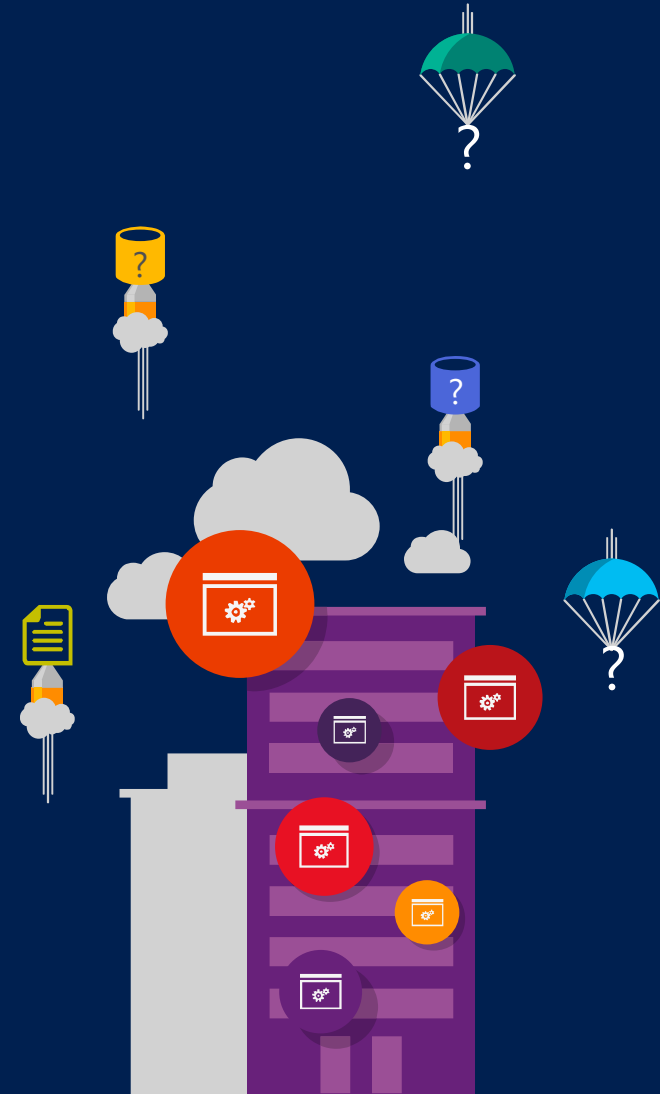
What will happen?

Predictive  
Analytics

How can we make it happen?

Prescriptive  
Analytics

# Data lake defined



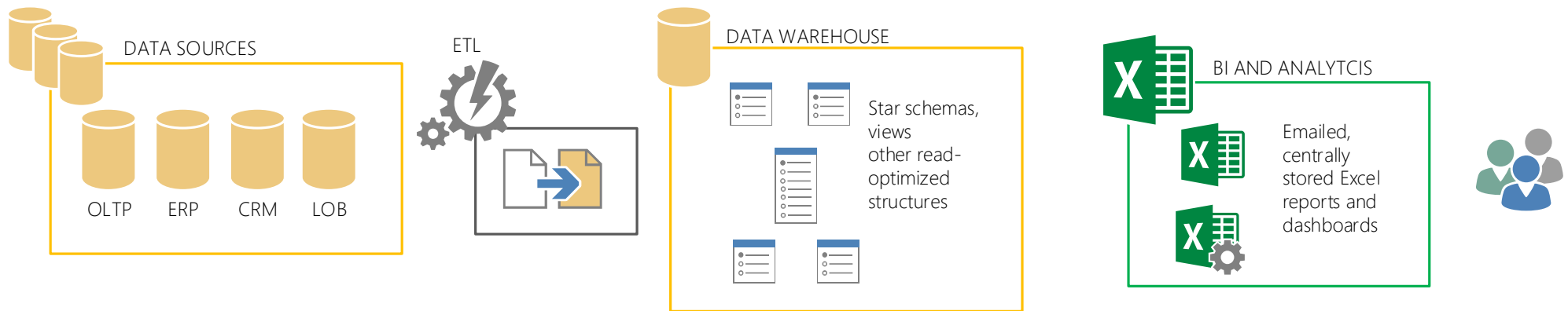
# What is a data lake?

A storage repository, usually Hadoop, that holds a vast amount of raw data in its native format until it is needed.

- A place to store unlimited amounts of data in any format **inexpensively**, especially for **archive purposes**
- Allows **collection of data** that you may or may not use later: “just in case”
- A way to describe any large data pool in which the schema and data requirements are not defined until the data is queried: “just in time” or “**schema on read**”
- **Complements EDW** and can be seen as a data source for the EDW – capturing all data but only passing relevant data to the EDW
- **Frees up expensive EDW resources** (storage and processing), especially for data refinement
- Allows for data exploration to be performed without waiting for the EDW team to model and load the data (**quick user access**)
- Some processing is better done with **Hadoop tools** than ETL tools like SSIS
- Easily scalable

# Traditional Approaches

Current state of a data warehouse



Well manicured, often relational sources

Known and expected data volume and formats

Little to no change



Complex, rigid transformations

Required extensive monitoring

Transformed historical into read structures



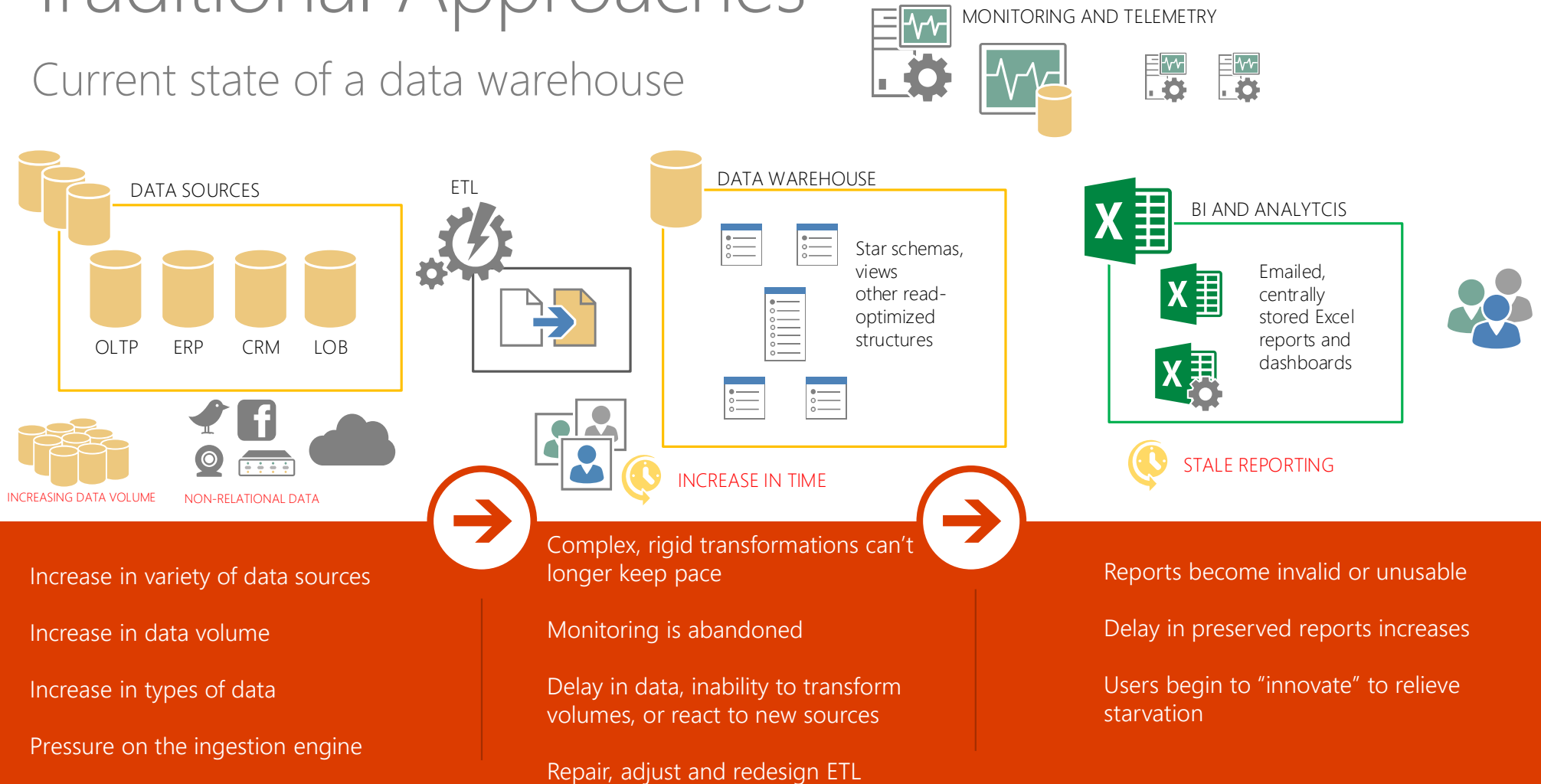
Flat, canned or multi-dimensional access to historical data

Many reports, multiple versions of the truth

24 to 48h delay

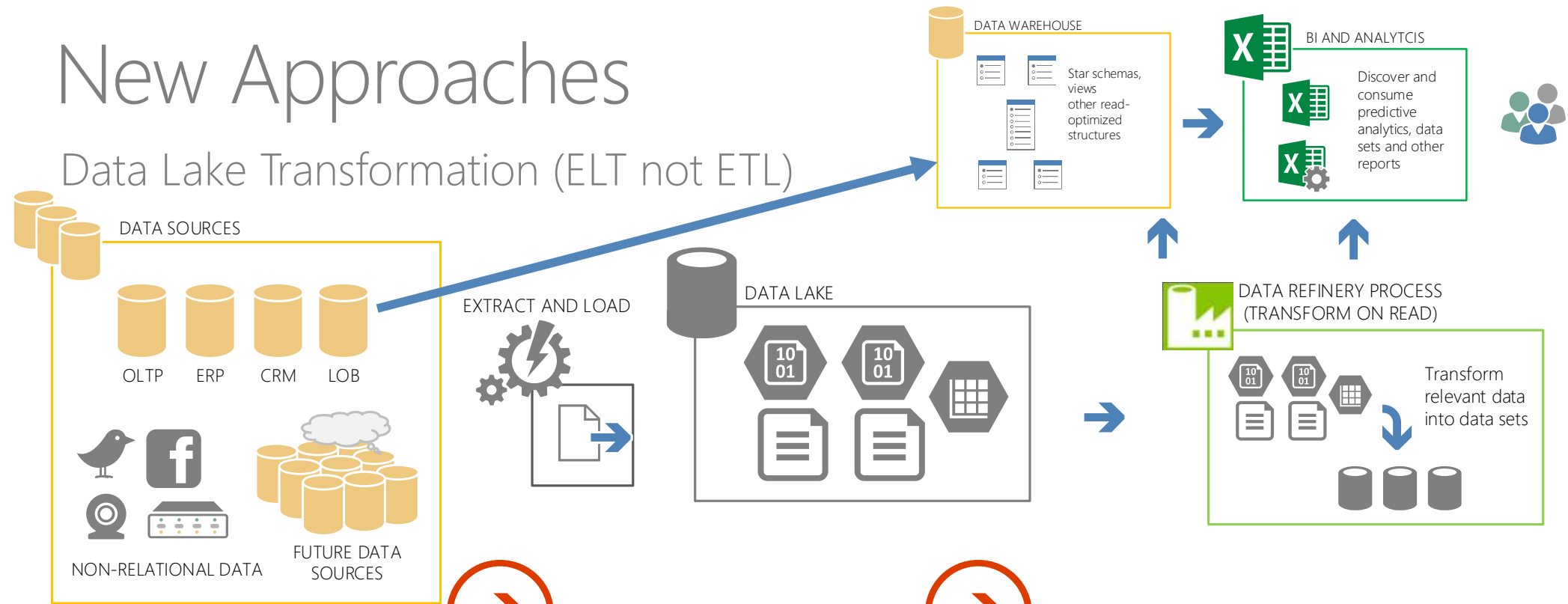
# Traditional Approaches

Current state of a data warehouse



# New Approaches

## Data Lake Transformation (ELT not ETL)



All data sources are considered

Leverages the power of on-prem technologies and the cloud for storage and capture

Native formats, streaming data, big data



Extract and load, no/minimal transform

Storage of data in near-native format

Orchestration becomes possible

Streaming data accommodation becomes possible



Refineries transform data on read

Produce curated data sets to integrate with traditional warehouses

Users discover published data sets/services using familiar tools

# Data Analysis Paradigm Shift

*OLD WAY: Structure -> Ingest -> Analyze*

*NEW WAY: Ingest -> Analyze -> Structure*

# 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!*



# Should I use Hadoop or NoSQL for the data lake?

*Most implementations use Hadoop as the data lake because of these benefits:*

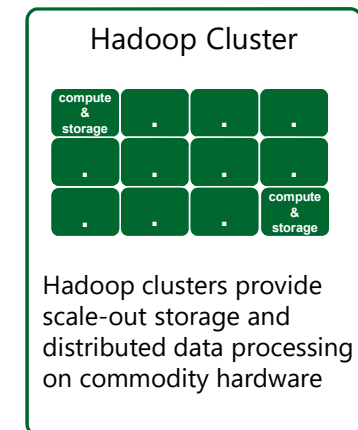
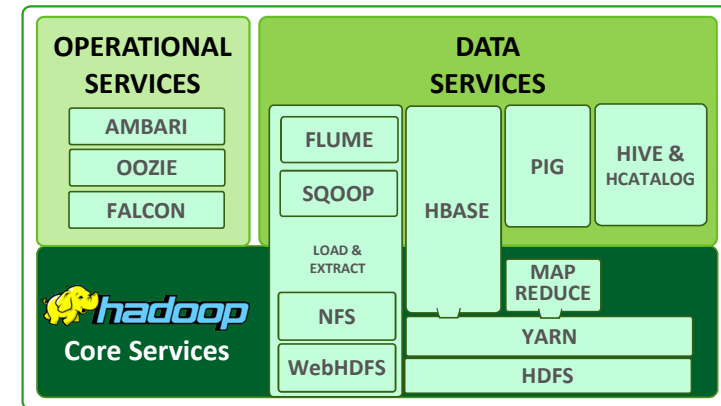
- Open-source software ecosystem that allows for massively parallel computing
- No inherent structure (no conversion to JSON needed)
- Good for batch processing, large files, volume writes, parallel scans, sequential access (NoSQL designed for large-scale OLTP)
- Large ecosystem of products
- Low cost
- Con: performance

# Hadoop as the data lake

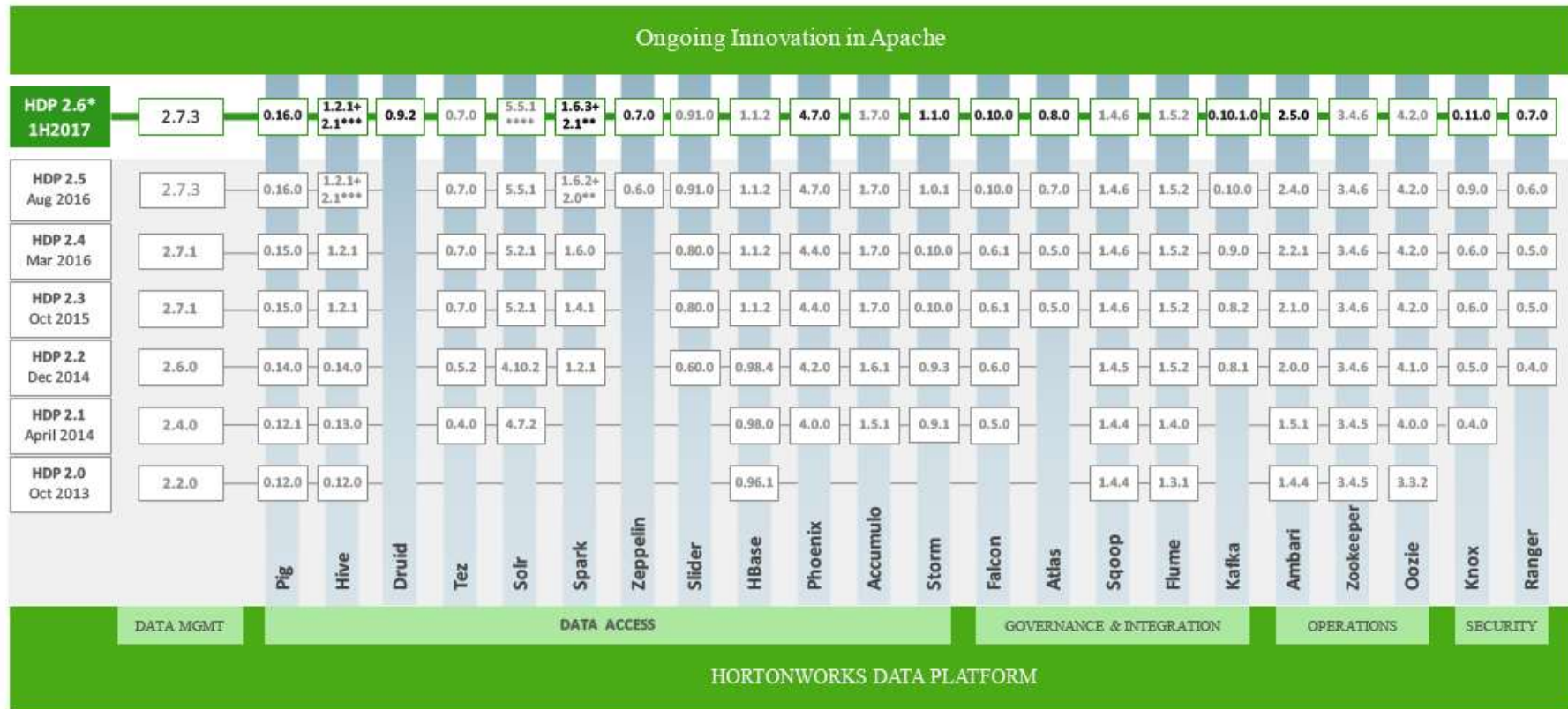


# What is Hadoop?

- Distributed, scalable system on commodity HW
- Composed of a few parts:
  - HDFS – Distributed file system
  - MapReduce – Programming model
  - Other tools: Hive, Pig, SQOOP, HCatalog, HBase, Flume, Mahout, YARN, Tez, Spark, Stinger, Oozie, ZooKeeper, Flume, Storm
- Main players are Hortonworks, Cloudera, MapR
- WARNING: Hadoop, while ideal for processing huge volumes of data, is inadequate for analyzing that data in real time (companies do batch analytics instead)



# Hortonworks Data Platform 2.5



\* HDP 2.6 – Shows current Apache branches being used. Final component version subject to change based on Apache release process.

\*\* Spark 1.6.3+ Spark 2.1 – HDP 2.6 supports both Spark 1.6.3 and Spark 2.1 as GA.

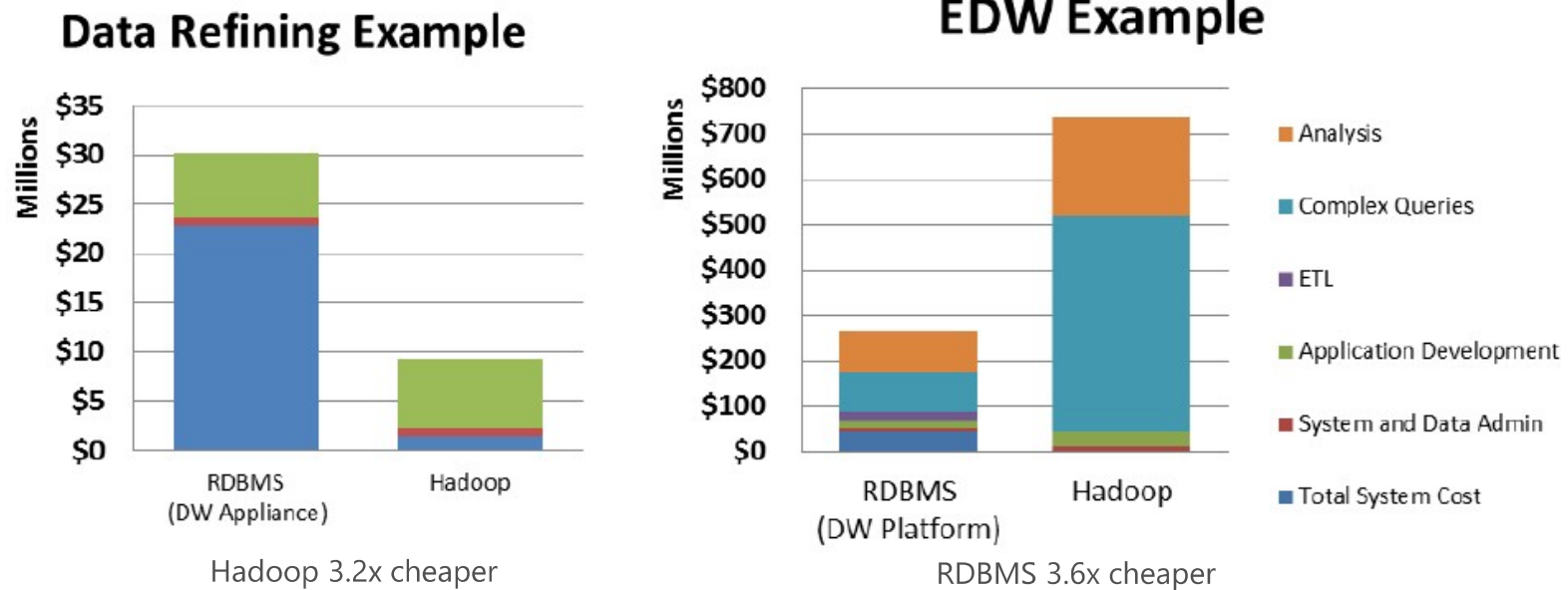
\*\*\* Hive 2.1 is GA within HDP 2.6.

\*\*\*\* Apache Solr is available as an add-on product HDP Search.

Simply put, Hortonworks ties all the open source products together (22)

# The real cost of Hadoop

Total solution cost (5 years)

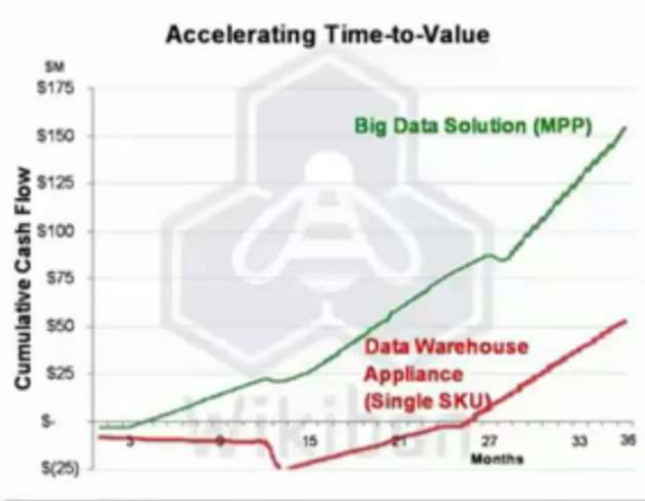


"Big Data – What Does It Really Cost?" Winter Corporation, 2013, <http://www.wintercorp.com/tcod-report/>

## INTRODUCTION: WILL THE EDW BE REPLACED BY HADOOP?



2011



Reference:  
[http://wikibon.org/wiki/Financial\\_Comparison\\_of\\_Big\\_Data\\_MPP\\_Solution\\_and\\_Data\\_Warehouse\\_Appliance](http://wikibon.org/wiki/Financial_Comparison_of_Big_Data_MPP_Solution_and_Data_Warehouse_Appliance)

Today

**Fewer than 5% of organizations actually plan to replace their data warehouse with Hadoop, and that percentage is dropping each year.**

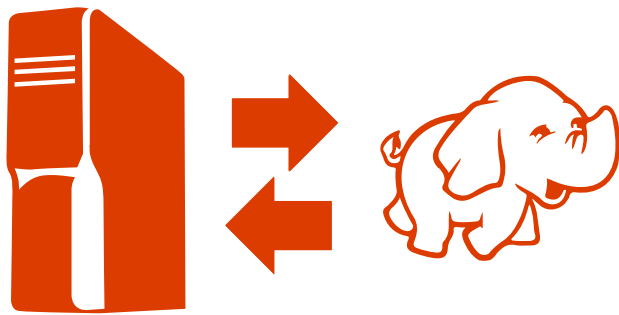
Reference: Gartner's Magic Quadrant for Data Warehouse Database Management Systems, November 2013

# Reasons not to use Hadoop as your DW

- Hadoop does not provide for very fast query reads. Dashboard user don't want to wait 10+ seconds for a MapReduce job to start up to execute a Hive query
- Hadoop lacks a query optimizer and indexing and performs poorly for complex queries
- Hadoop is not relational, as all the data is in files in HDFS, so there always is a conversion process to convert the data to a relational format
- Hadoop is not a database management system. It does not have functionality such as update of data, referential integrity, statistics, ACID compliance, data security, and the plethora of tools and facilities needed to govern corporate data assets
- Restricted SQL support, such as certain aggregate functions missing
- There is no metadata stored in HDFS, so another tool needs to be used to store that, adding complexity and slowing performance
- Finding expertise in Hadoop is very difficult: The small number of people who understand Hadoop and all its various versions and products versus the large number of people who know SQL
- Super complex, lot's of integration with multiple technologies to make it work
- Many tools/technologies/versions/vendors (fragmentation), no standards, difficult to make a corporate standard
- Some reporting tools don't work against Hadoop, as well as some reporting tools require data in OLAP
- May require end-users to learn new reporting tools and Hadoop technologies to query the data
- The new Hadoop solutions (Tez, X, Spark, etc) are still figuring themselves out. Customers should not take the risk of investing in one of these solutions (like MapReduce) that may be obsolete
- It might not save you much in costs: you still have to purchase hardware, support, licenses, training, migration costs. And then there is the possibility of a big company acquiring a Hadoop company and the cheap solution becoming much more expensive
- If you need to combine relational data with Hadoop, you will need to move that relational data to Hadoop since there is no PolyBase-like technology

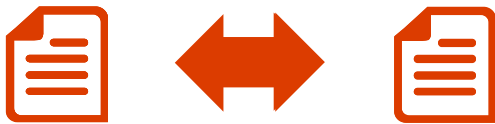
# Use cases using Hadoop and a DW in combination

Bringing islands of Hadoop data together



Archiving data warehouse data to Hadoop (move)  
(Hadoop as cold storage)

Exporting relational data to Hadoop (copy)  
(Hadoop as backup/DR, analysis, cloud use)



Importing Hadoop data into data warehouse (copy)  
(Hadoop as staging area, sandbox, Data Lake)



# Modern Data Warehouse



# Modern Data Warehouse

Think about future needs:

- Increasing data volumes
- Real-time performance
- New data sources and types
- Cloud-born data
- Multi-platform solution
- Hybrid architecture

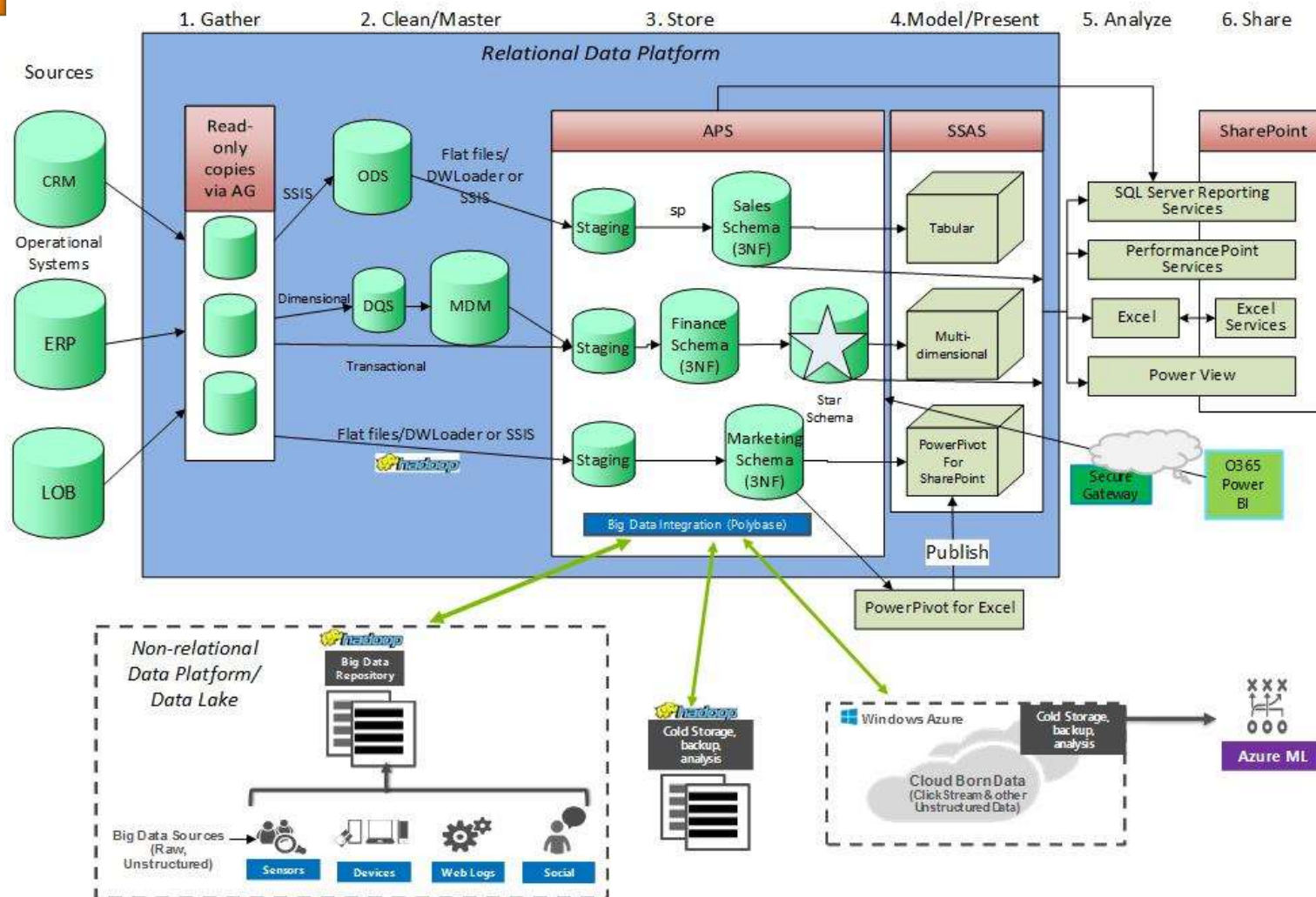
The  
Dream

## Modern Data Warehouse

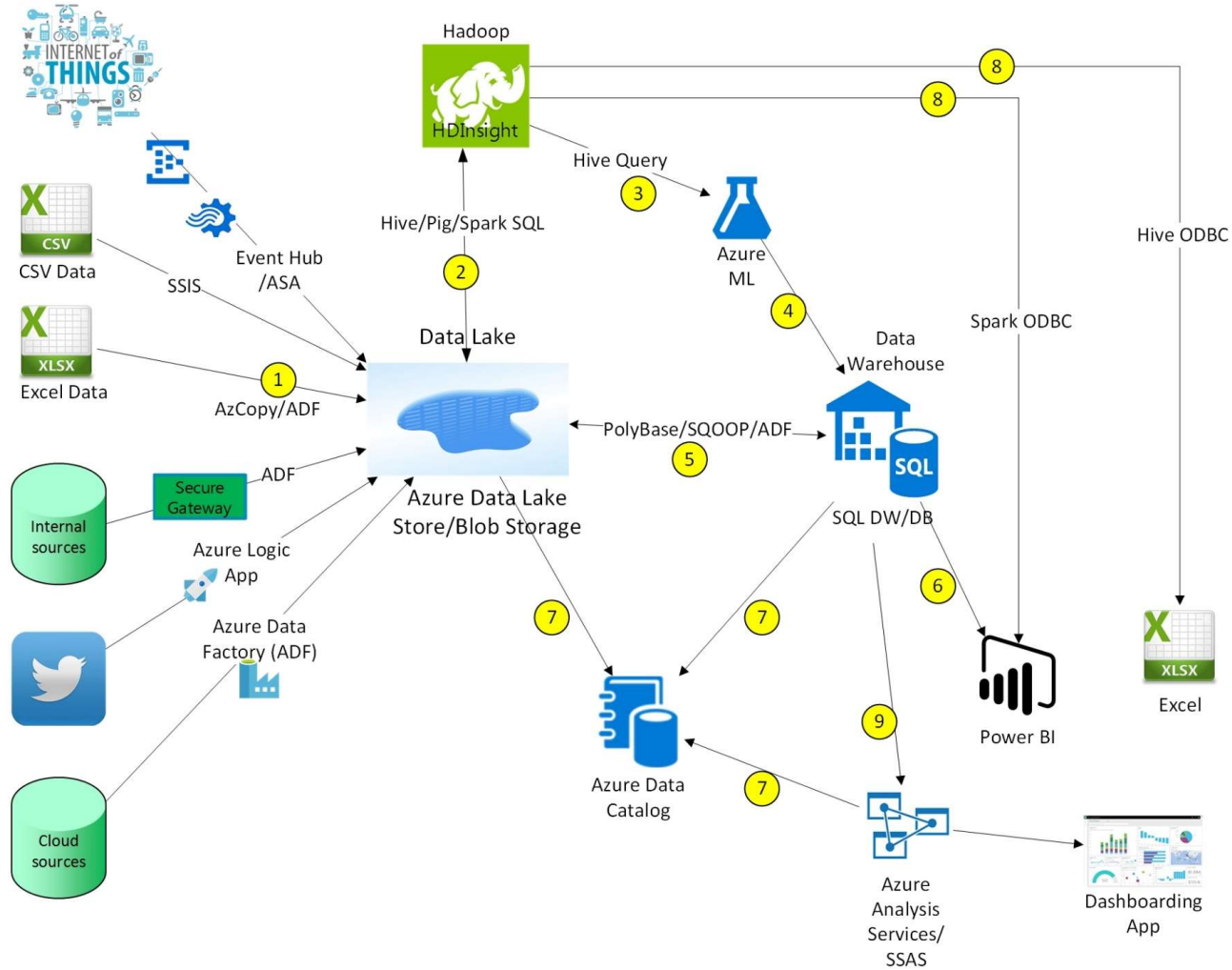


The  
Reality

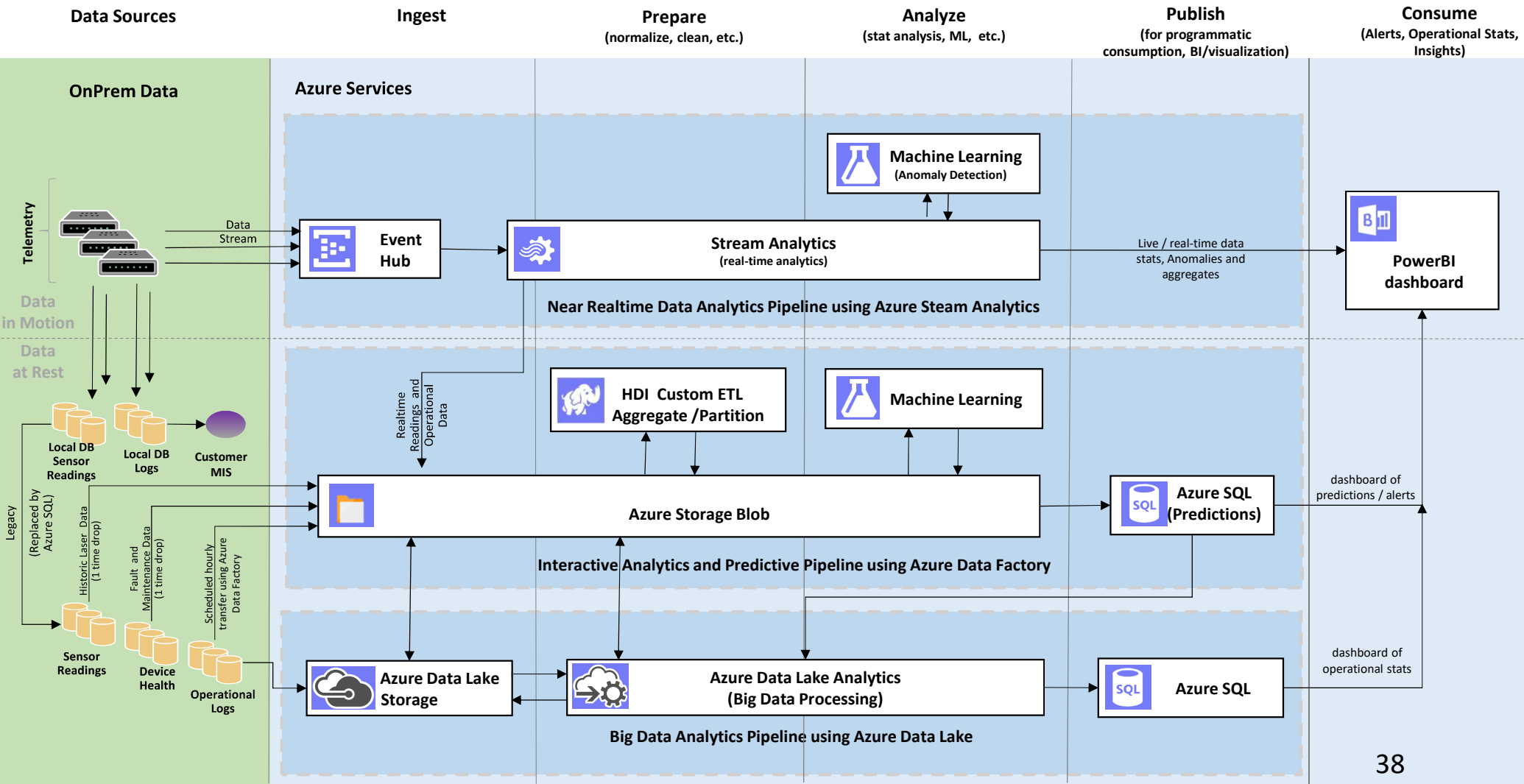
## Modern Data Warehouse



# The Reality



Base Architecture : Big Data Advanced Analytics Pipeline



# Roles when using both Data Lake and DW

## Data Lake/Hadoop (staging and processing environment)

- Batch reporting
- Data refinement/cleaning
- ETL workloads
- Store historical data
- Sandbox for data exploration
- One-time reports
- Data scientist workloads
- Quick results

## Data Warehouse/RDBMS (serving and compliance environment)

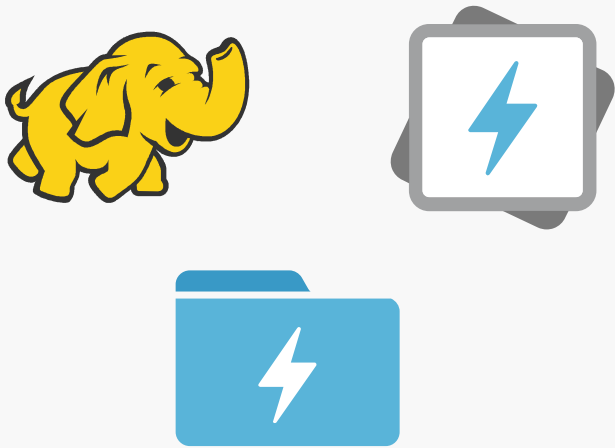
- Low latency
- High number of users
- Additional security
- Large support for tools
- Easily create reports (Self-service BI)
- *A data lake is just a glorified file folder with data files in it – how many end-users can accurately create reports from it?*

# Comparing a Data Lake and a Data Warehouse

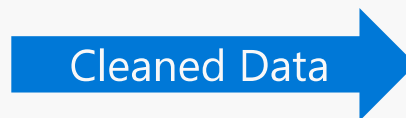
Data Lake	Data Warehouse
Complementary to DW	Can be sourced from Data Lake
System of Insight	System of Record
Schema-on-read	Schema-on-write
Detailed Data	Refined Data
Optimized for Cost	Optimized for I/O and CPU
Data Discovery	Data Reusability
Low User Concurrency	High User Concurrency
Varying Query Perf	Predictable Query Perf



# Complementing Each Other



Ingest any data  
Cleanse data  
Catalog data  
AA & ML



Structured Analysis  
Supports Interactive BI  
Quick Aggregations  
Applied Business Rules

# Microsoft data platform solutions

Product	Category	Description	More Info
SQL Server 2016	RDBMS	Earned top spot in Gartner's Operational Database magic quadrant. <b>JSON support</b>	<a href="https://www.microsoft.com/en-us/server-cloud/products/sql-server-2016/">https://www.microsoft.com/en-us/server-cloud/products/sql-server-2016/</a>
SQL Database	RDBMS/DBaaS	Cloud-based service that is provisioned and scaled quickly. Has built-in high availability and disaster recovery. <b>JSON support</b>	<a href="https://azure.microsoft.com/en-us/services/sql-database/">https://azure.microsoft.com/en-us/services/sql-database/</a>
SQL Data Warehouse	MPP RDBMS/DBaaS	Cloud-based service that handles relational big data. Provision and scale quickly. Can pause service to reduce cost	<a href="https://azure.microsoft.com/en-us/services/sql-data-warehouse/">https://azure.microsoft.com/en-us/services/sql-data-warehouse/</a>
Analytics Platform System (APS)	MPP RDBMS	Big data analytics appliance for high performance and seamless integration of all your data	<a href="https://www.microsoft.com/en-us/server-cloud/products/analytics-platform-system/">https://www.microsoft.com/en-us/server-cloud/products/analytics-platform-system/</a>
Azure Data Lake Store	Hadoop storage	Removes the complexities of ingesting and storing all of your data while making it faster to get up and running with batch, streaming, and interactive analytics	<a href="https://azure.microsoft.com/en-us/services/data-lake-store/">https://azure.microsoft.com/en-us/services/data-lake-store/</a>
Azure Data Lake Analytics	On-demand analytics job service/Big Data-as-a-service	Cloud-based service that dynamically provisions resources so you can run queries on exabytes of data. Includes U-SQL, a new big data query language	<a href="https://azure.microsoft.com/en-us/services/data-lake-analytics/">https://azure.microsoft.com/en-us/services/data-lake-analytics/</a>
HDInsight	PaaS Hadoop compute	A managed Apache Hadoop, Spark, R, HBase, and Storm cloud service made easy	<a href="https://azure.microsoft.com/en-us/services/hdinsight/">https://azure.microsoft.com/en-us/services/hdinsight/</a>
DocumentDB	PaaS NoSQL: Document Store	Get your apps up and running in hours with a fully managed NoSQL database service that indexes, stores, and queries data using familiar SQL syntax	<a href="https://azure.microsoft.com/en-us/services/documentdb/">https://azure.microsoft.com/en-us/services/documentdb/</a>
Azure Table Storage	PaaS NoSQL: Key-value Store	Store large amount of semi-structured data in the cloud	<a href="https://azure.microsoft.com/en-us/services/storage/tables/">https://azure.microsoft.com/en-us/services/storage/tables/</a>

# Cortana Intelligence Suite

Integrated as part of an end-to-end suite

