Big data architectures and the data lake

James Serra
Big Data Evangelist
Microsoft
JamesSerra3@gmail.com

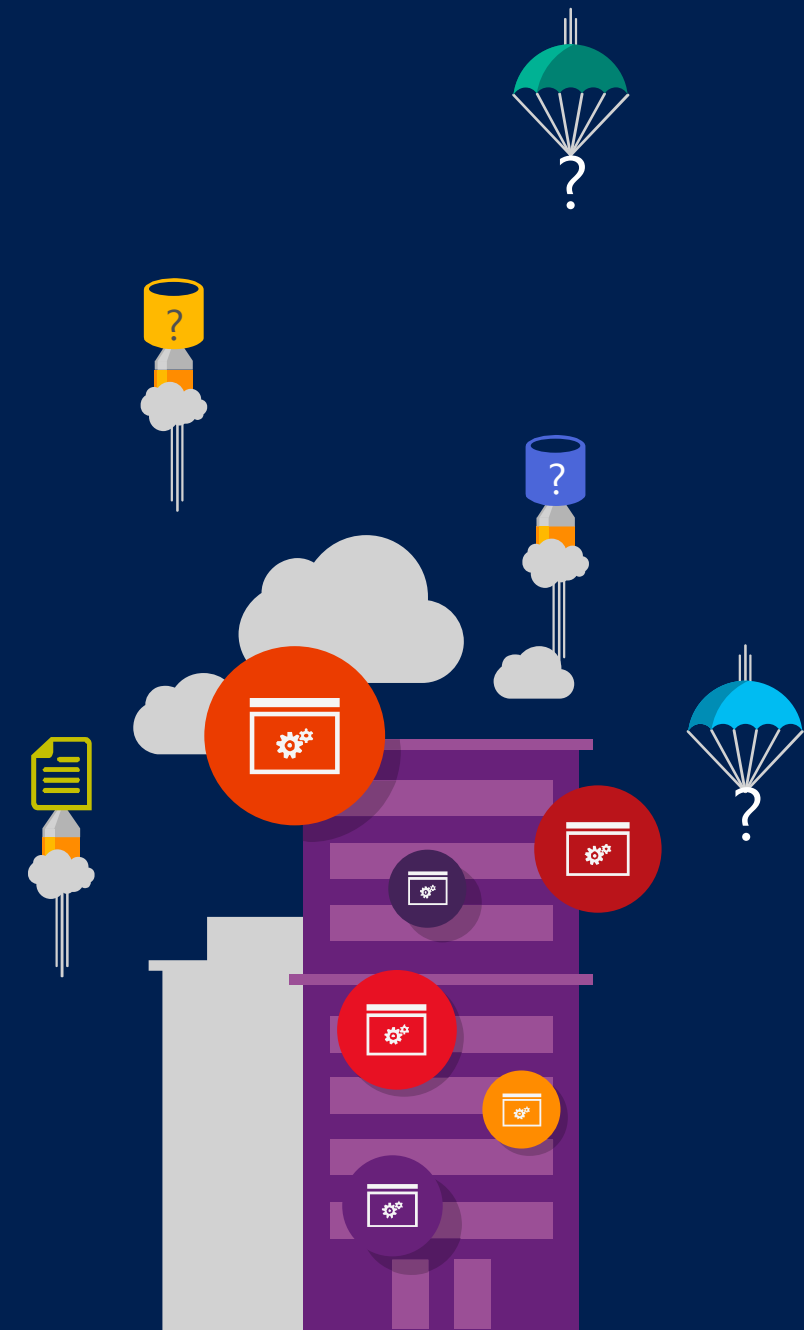
About Me

- Microsoft, Big Data Evangelist
- In IT for 30 years, worked on many BI and DW projects
- Worked as desktop/web/database developer, DBA, BI and DW architect and developer, MDM architect, PDW/APS developer
- Been perm employee, contractor, consultant, business owner
- Presenter at PASS Business Analytics Conference, PASS Summit, Enterprise Data World conference
- Certifications: MCSE: Data Platform, Business Intelligence; MS: Architecting Microsoft Azure Solutions, Design and Implement Big Data Analytics Solutions, Design and Implement Cloud Data Platform Solutions
- Blog at JamesSerra.com
- Former SQL Server MVP
- Author of book "Reporting with Microsoft SQL Server 2012"

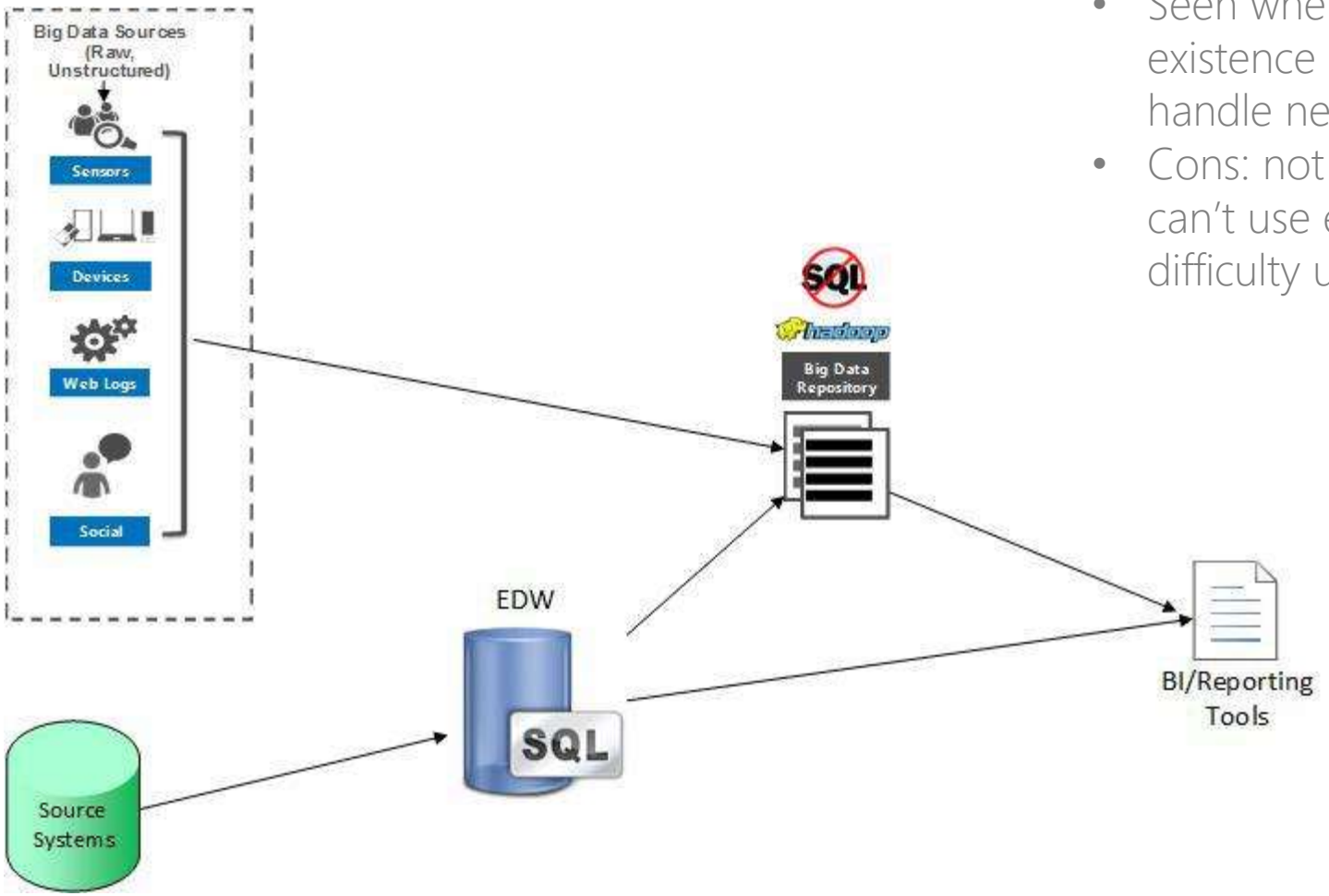
Agenda

- Big Data Architectures
- Why data lakes?
- Top-down vs Bottom-up
- Data lake defined
- Hadoop as the data lake
- Modern Data Warehouse
- Federated Querying
- Solution in the cloud
- SMP vs MPP

Big Data Architectures

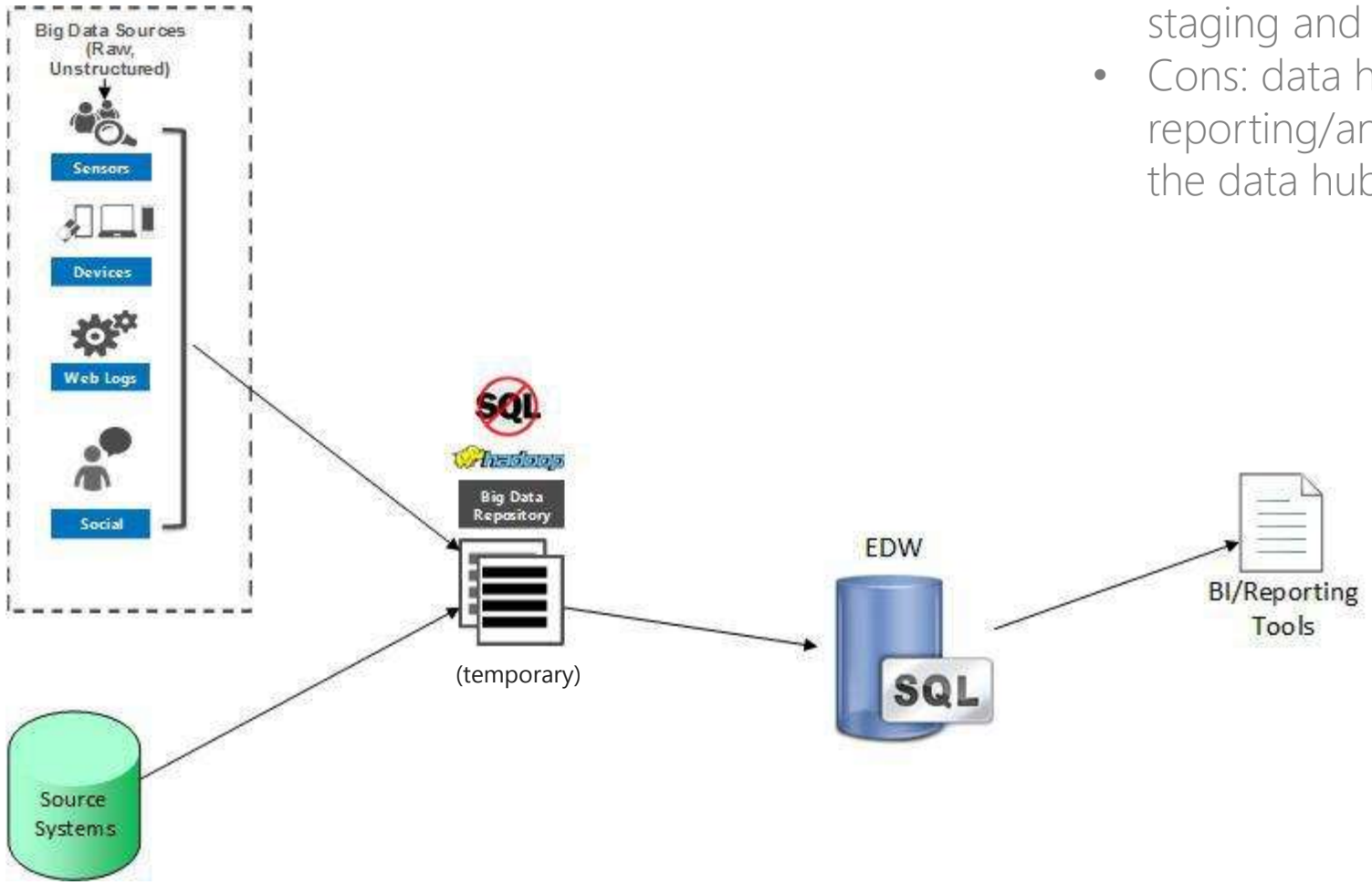


Enterprise data warehouse augmentation



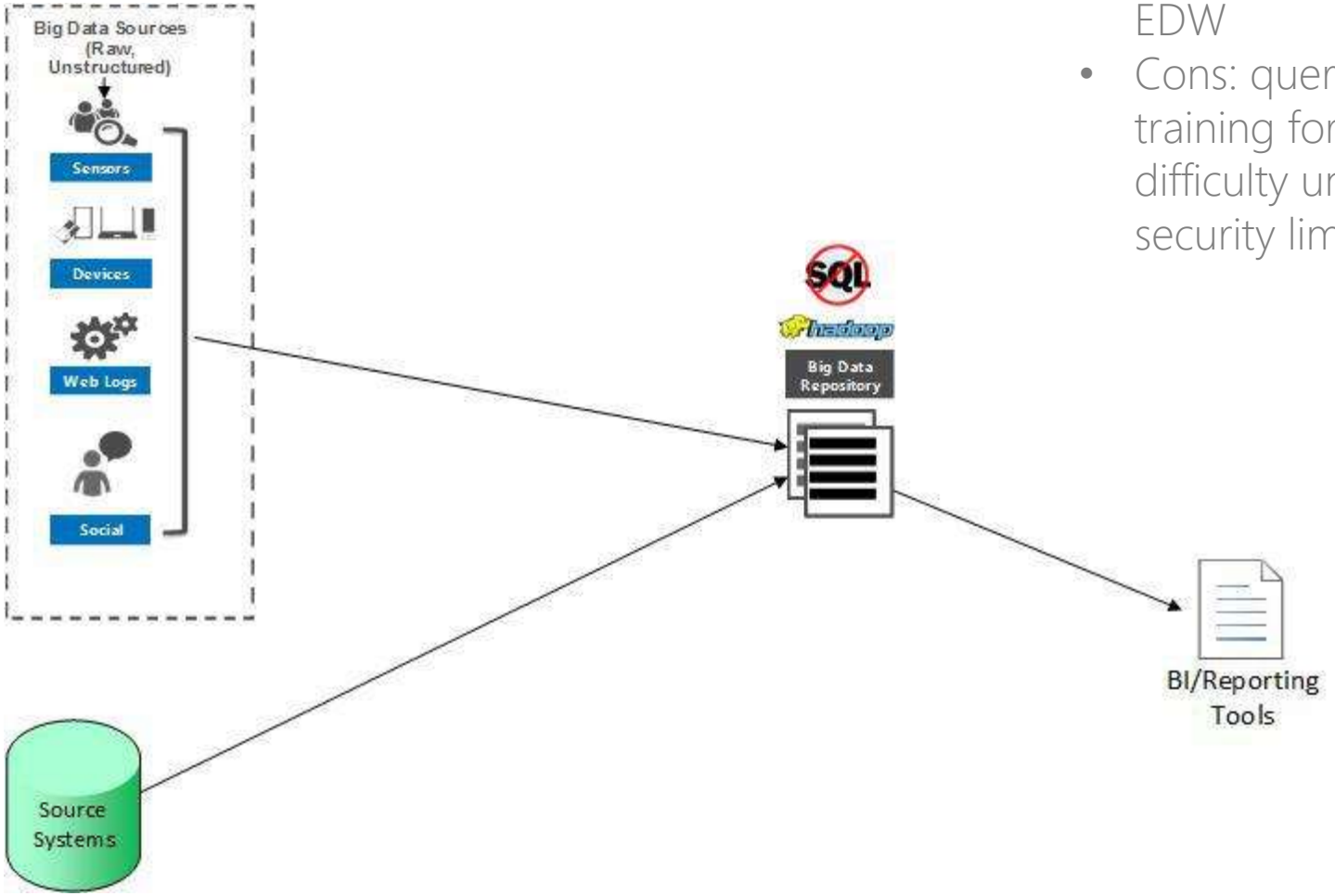
- Seen when EDW has been in existence a while and EDW can't handle new data
- Cons: not offloading EDW work, can't use existing tools, data hub difficulty understanding data

Data hub plus EDW



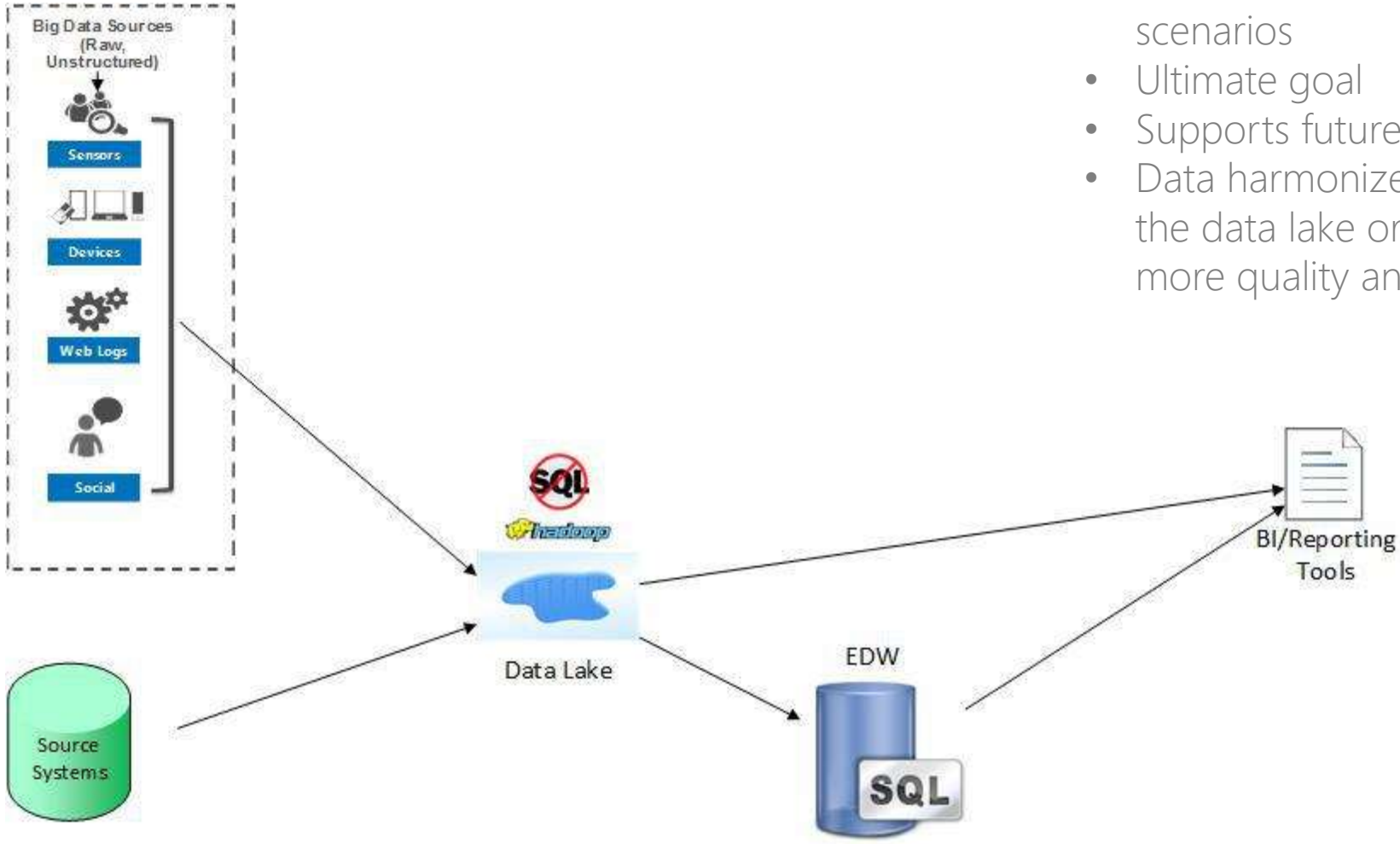
- Data hub is used as temporary staging and refining, no reporting
- Cons: data hub is temporary, no reporting/analyzing done with the data hub

All-in-one



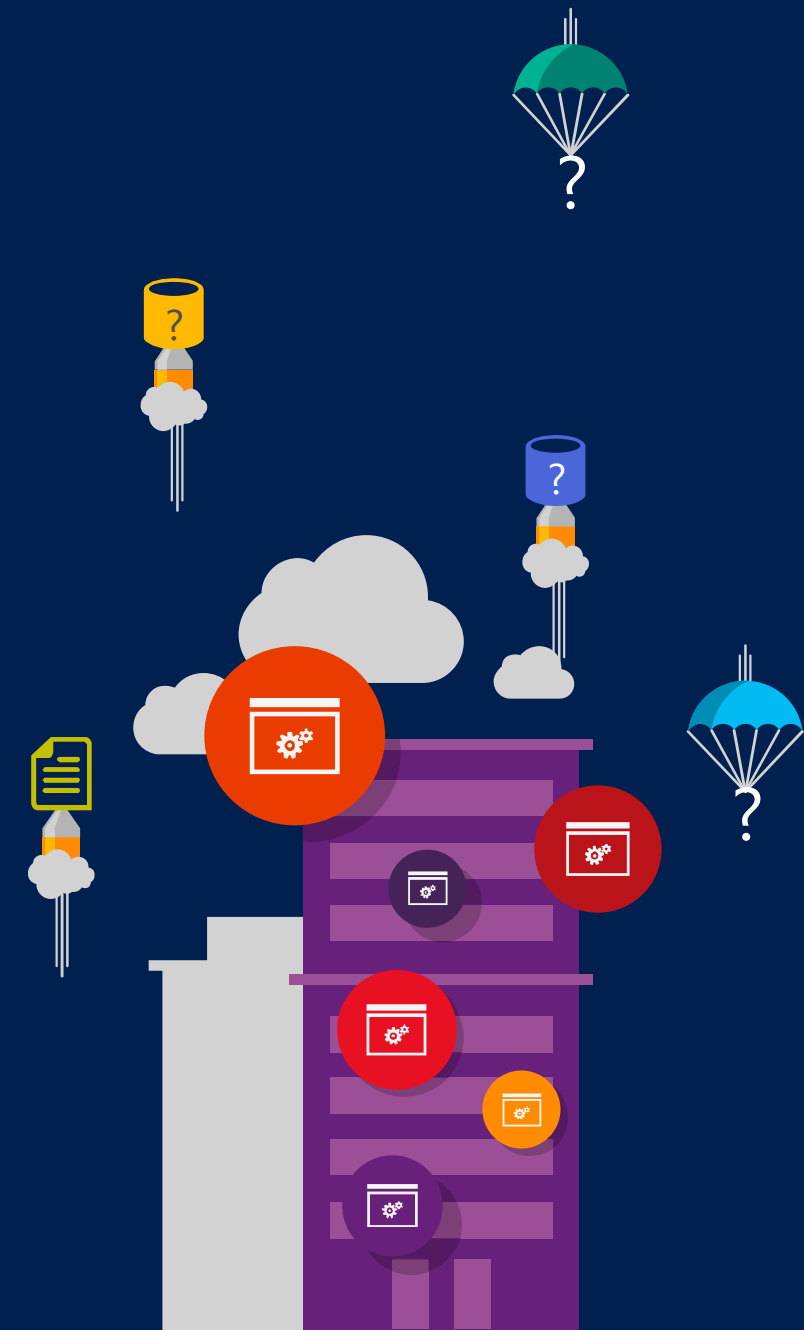
- Data hub is total solution, no EDW
- Cons: queries are slower, new training for reporting tools, difficulty understanding data, security limitations

Modern Data Warehouse



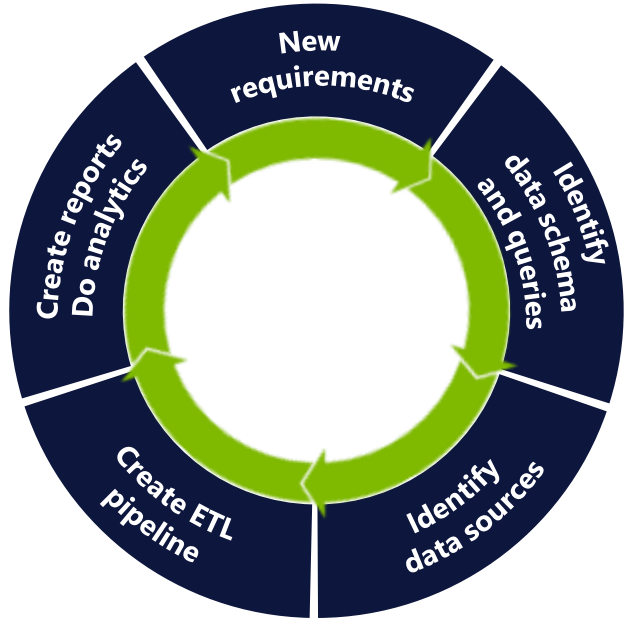
- Evolution of three previous scenarios
- Ultimate goal
- Supports future data needs
- Data harmonized and analyzed in the data lake or moved to EDW for more quality and performance

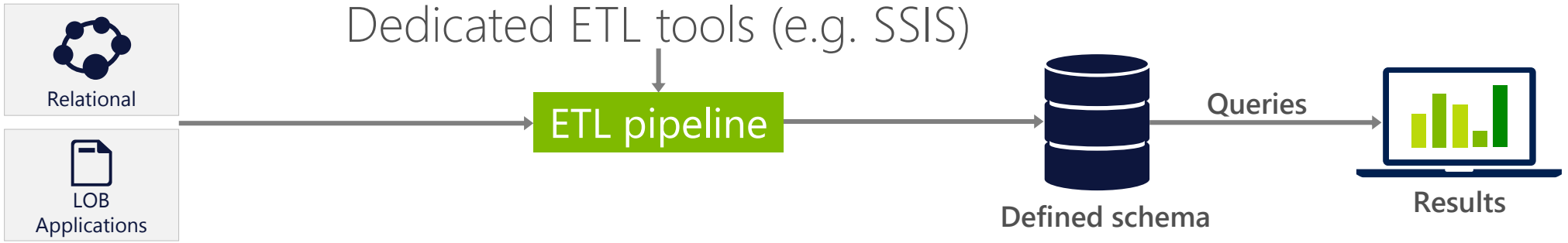
Why data lakes?



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





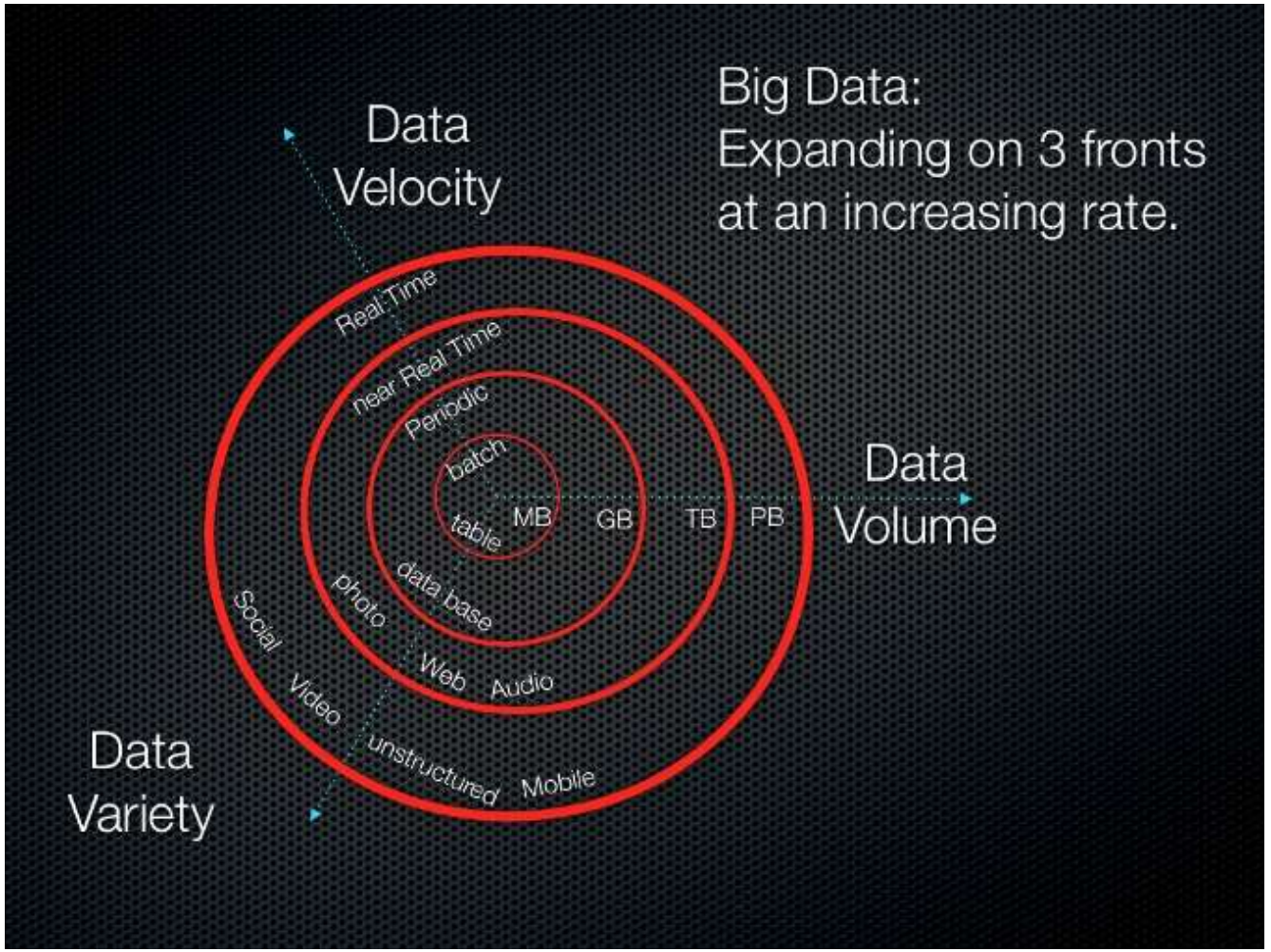
All data not immediately required is discarded or archived

Need to collect any data

Harness the growing and changing nature of data

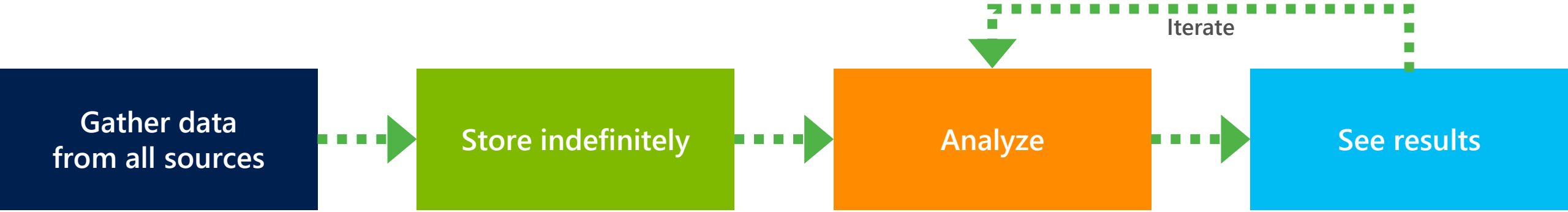
- Challenge is combining transactional data stored in relational databases with less structured data
- Big Data = All Data
- Get the right information to the right people at the right time in the right format

The three V's


Big Data:
Expanding on 3 fronts
at an increasing rate.
Data
Velocity
Real Time
near Real Time
Periodic
batch
table
MB
GB
TB
PB
Data
Volume
data base
photo
Social
Web
Audio
Video
unstructured
Mobile
Data
Variety

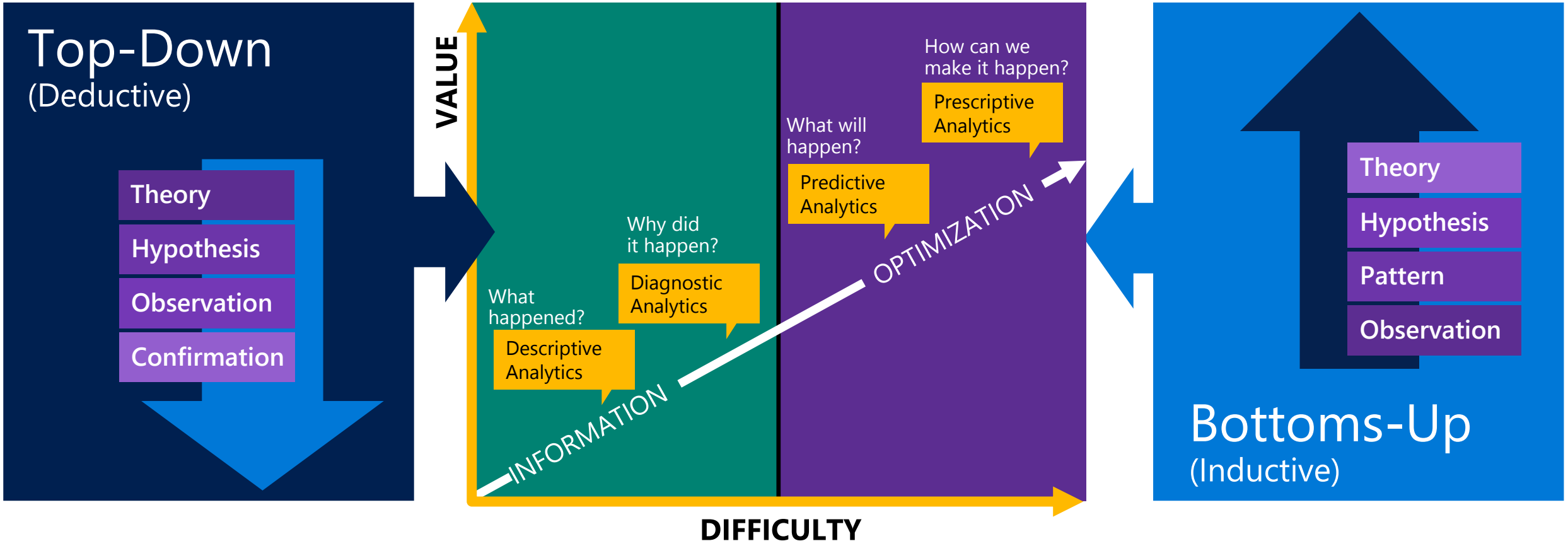
New big data thinking: All data has value

- 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



Top-down vs Bottom-up

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



Data Warehousing Uses A Top-Down Approach

Understand Corporate Strategy

Gather Requirements

Business Requirements

Technical Requirements

Implement Data Warehouse

- Reporting & Analytics Design → Reporting & Analytics Development
- Dimension Modelling → Physical Design
- ETL Design → ETL Development
- Setup Infrastructure → Install and Tune

BI and analytic

Dashboards Reporting

Data warehouse

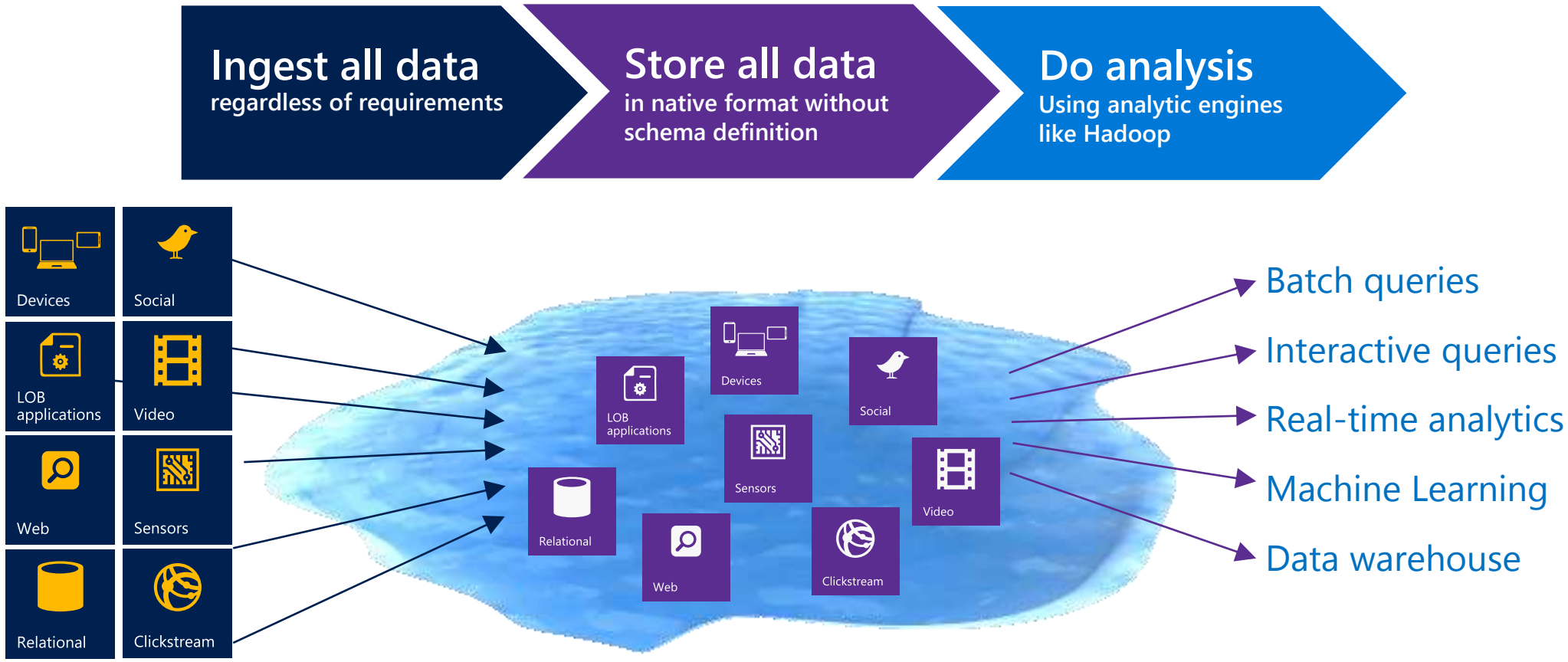
01100
10010
00101

ETL

Data sources

OLTP ERP CRM LOB

The “data lake” Uses A Bottoms-Up Approach



Data Lake + Data Warehouse Better Together

What happened?

Descriptive Analytics

Why did it happen?

Diagnostic Analytics

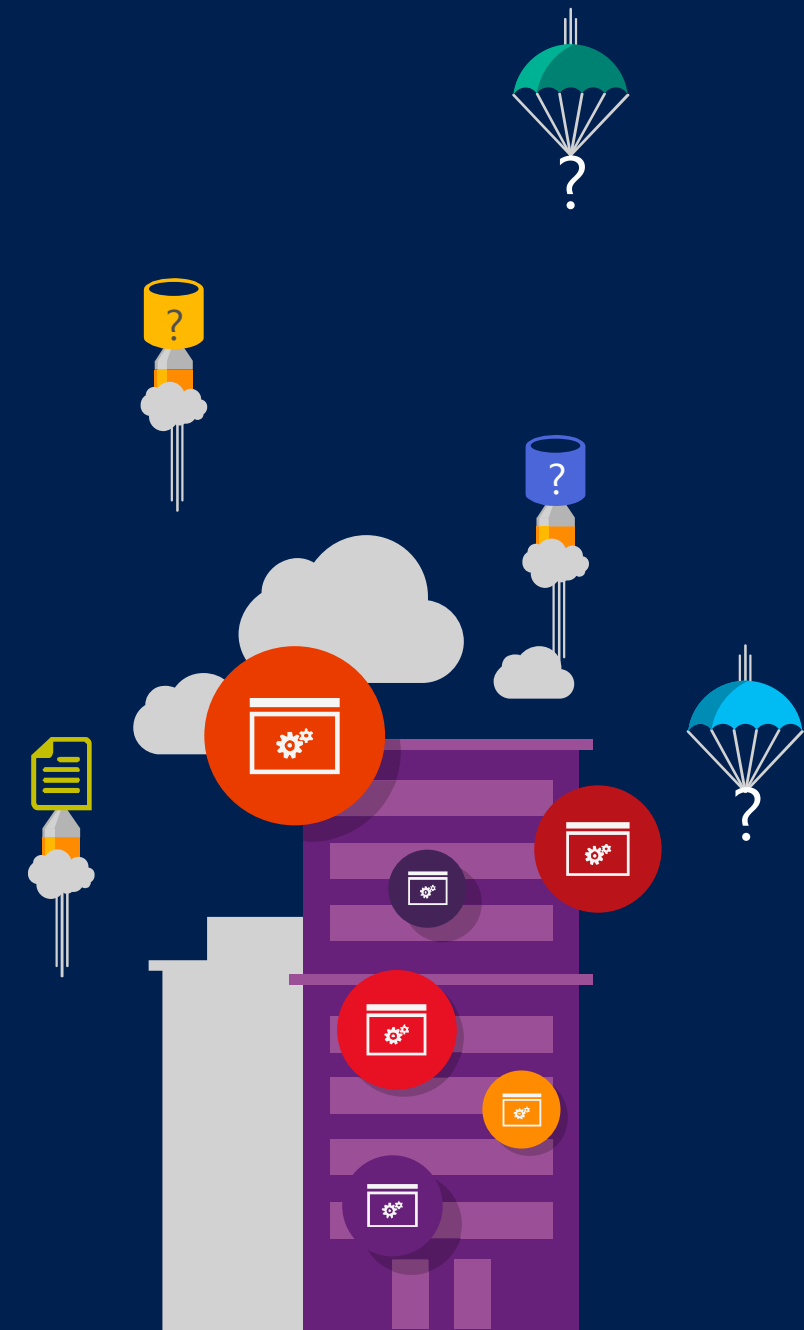
What will happen?

Predictive Analytics

How can we make it happen

Prescriptive Analytics

Data lake defined



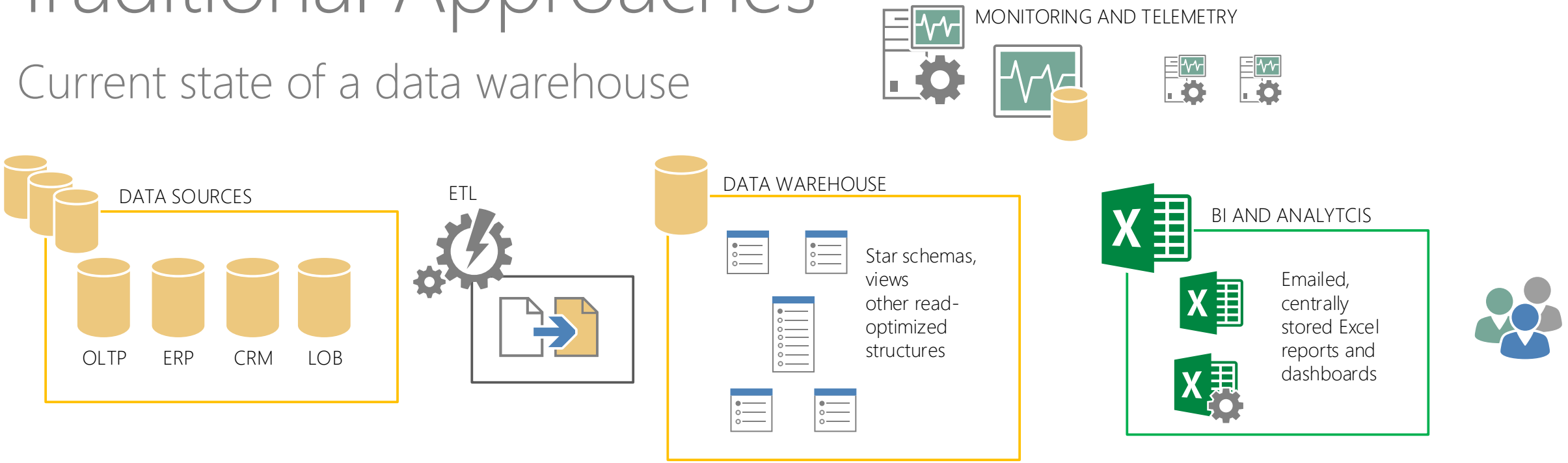
What is a data lake?

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 in 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

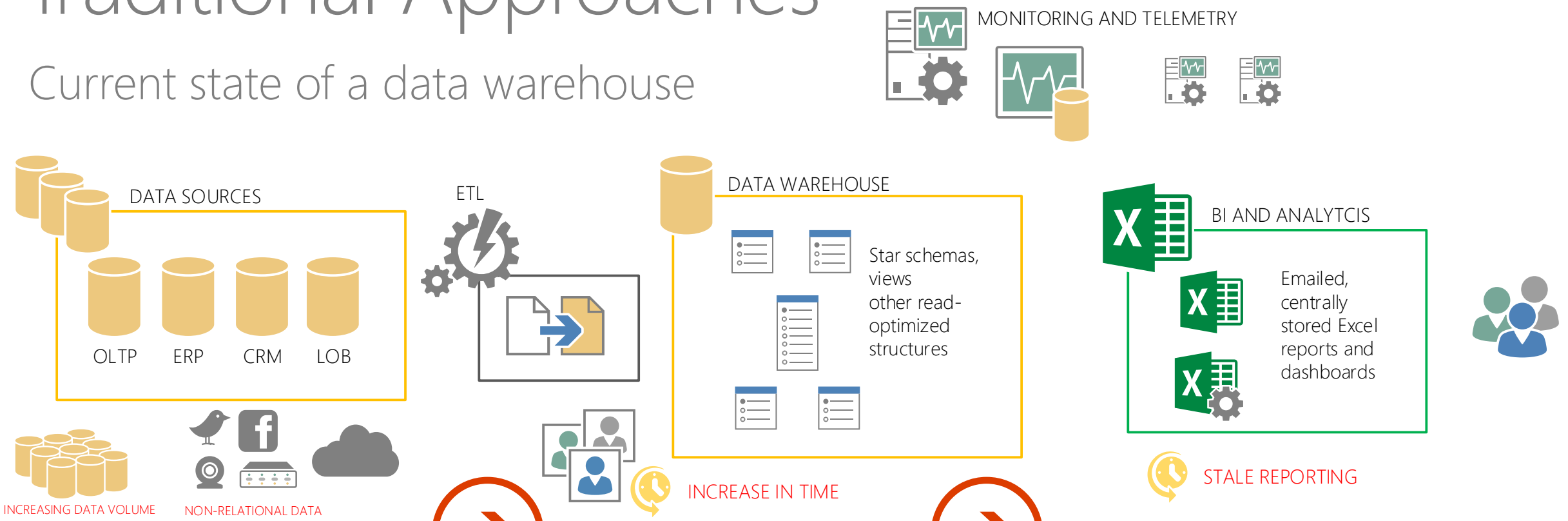
Flat, canned or multi-dimensional access to historical data

Many reports, multiple versions of the truth

24 to 48h delay

Traditional Approaches

Current state of a data warehouse



Increase in variety of data sources

Increase in data volume

Increase in types of data

Pressure on the ingestion engine

Complex, rigid transformations can't longer keep pace

Monitoring is abandoned

Delay in data, inability to transform volumes, or react to new sources

Repair, adjust and redesign ETL

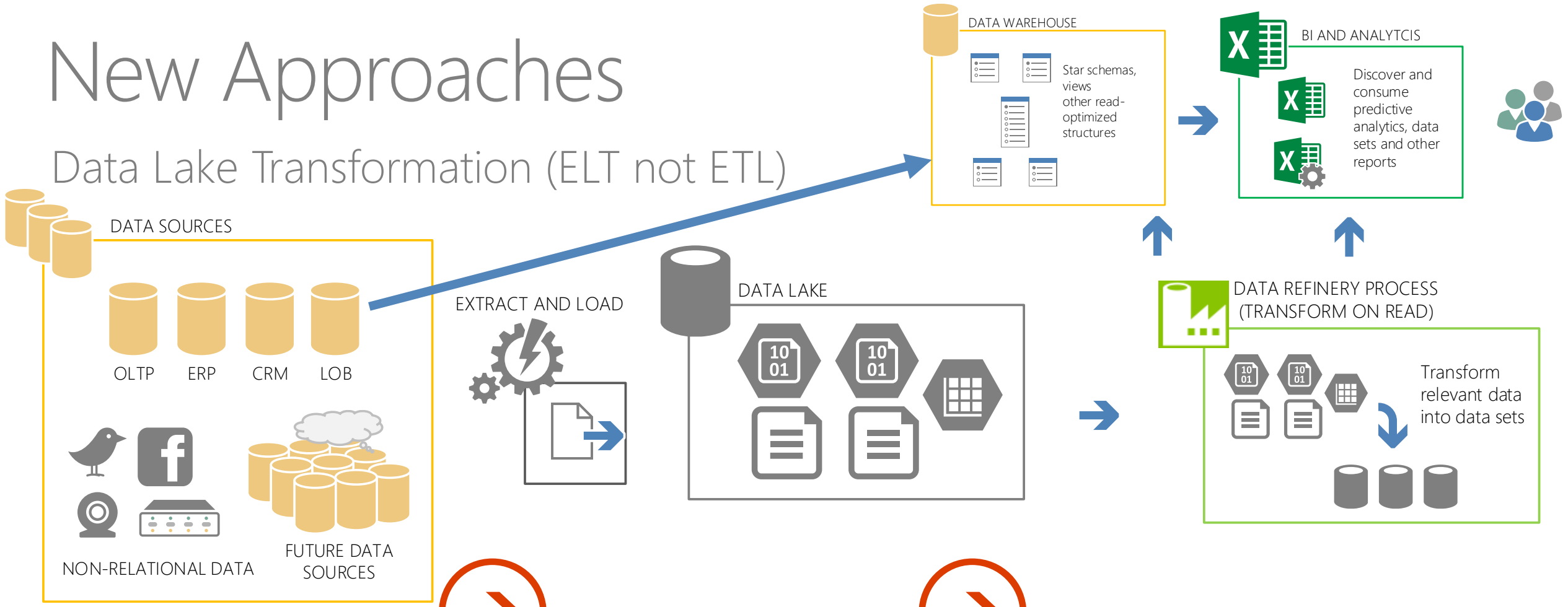
Reports become invalid or unusable

Delay in preserved reports increases

Users begin to "innovate" to relieve starvation

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?

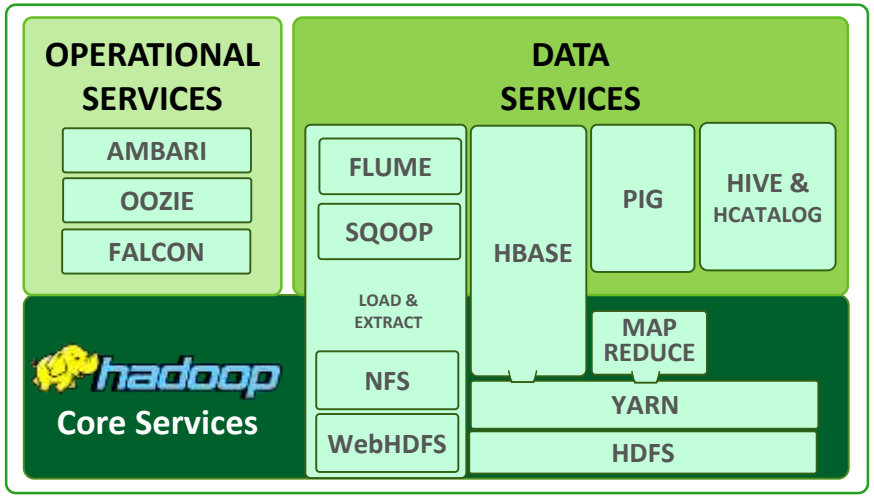
Most implementations use Hadoop as the data lake because of these benefits:

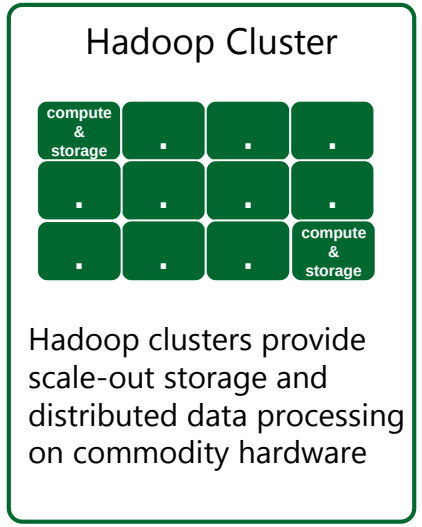
- Open-source software ecosystem that allows for massively parallel computing
- No inherent structure (no conversion to JSON needed)
- Good for batch processing, large files, volume writes, parallel scans, sequential access (NoSQL designed for large-scale OLTP)
- Large ecosystem of products
- Low cost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

Hadoop as the data lake

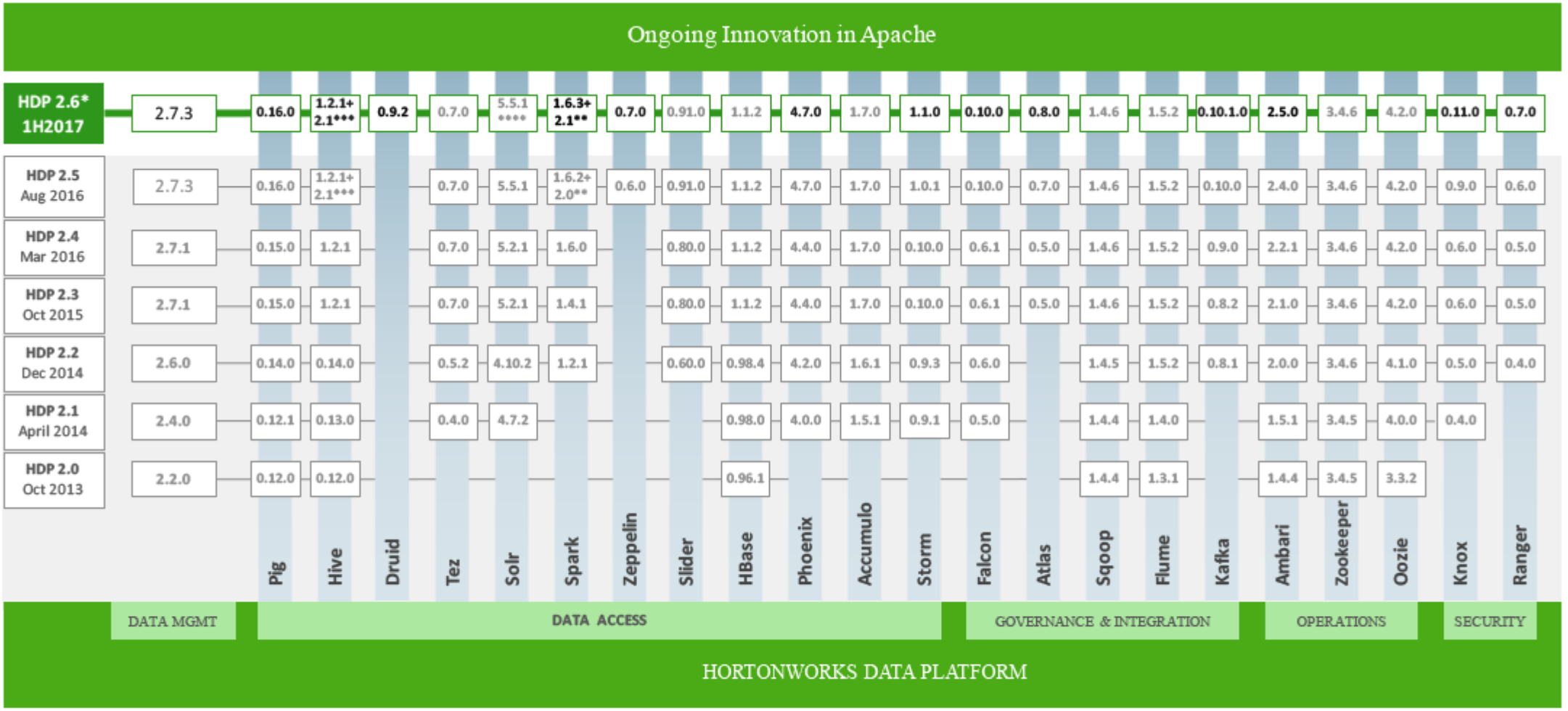
What is Hadoop?

- 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

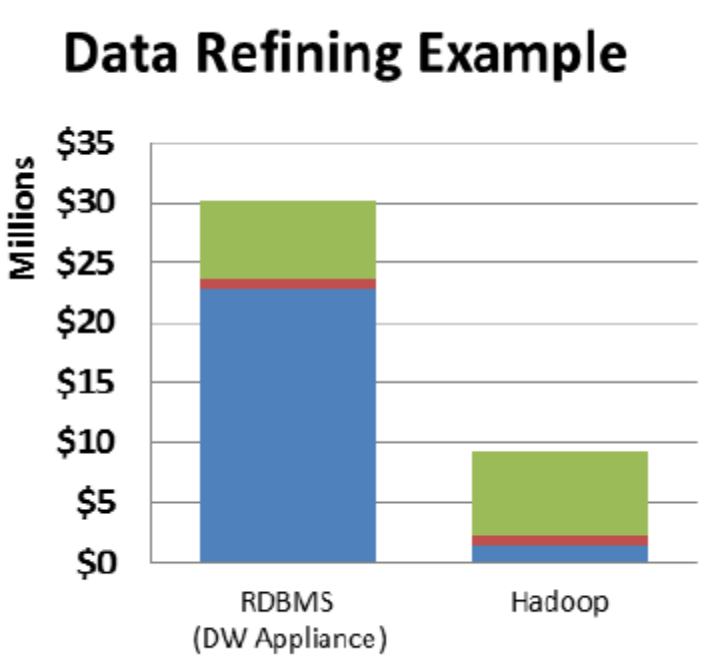


** HDP 2.6 – Shows current Apache branches being used. Final component version subject to change based on Apache release process.*
*** Spark 1.6.3+ Spark 2.1 – HDP 2.6 supports both Spark 1.6.3 and Spark 2.1 as GA.*
**** Hive 2.1 is GA within HDP 2.6.*
***** Apache Solr is available as an add-on product HDP Search.*

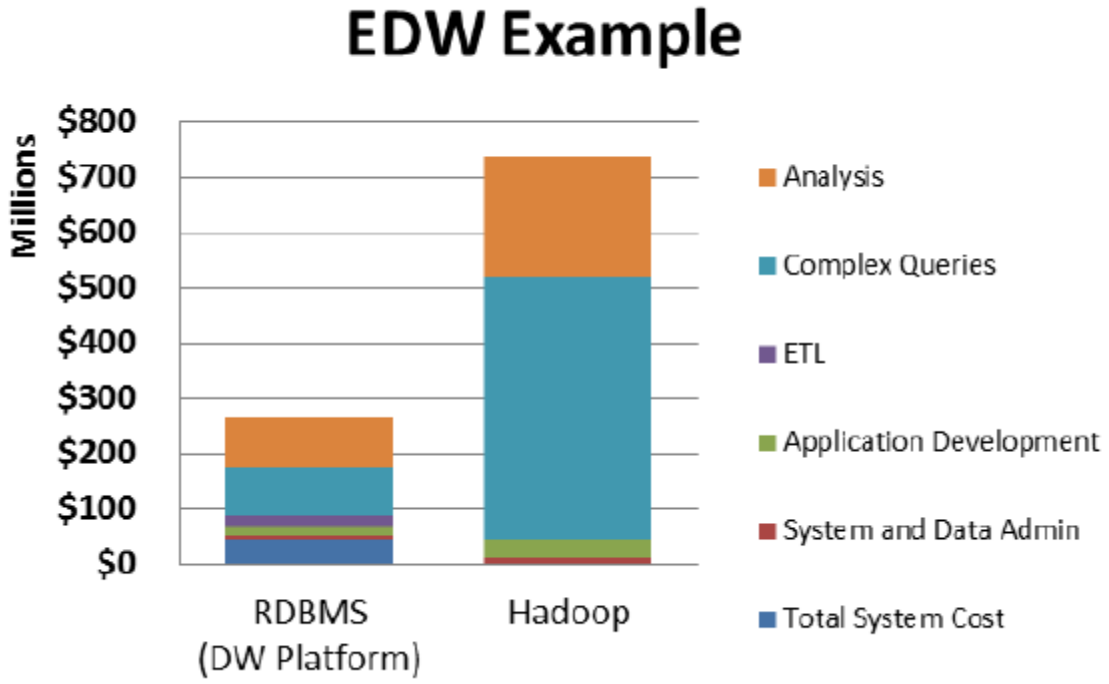
Simply put, Hortonworks ties all the open source products together (22)

The real cost of Hadoop

Total solution cost (5 years)



Hadoop 3.2x cheaper

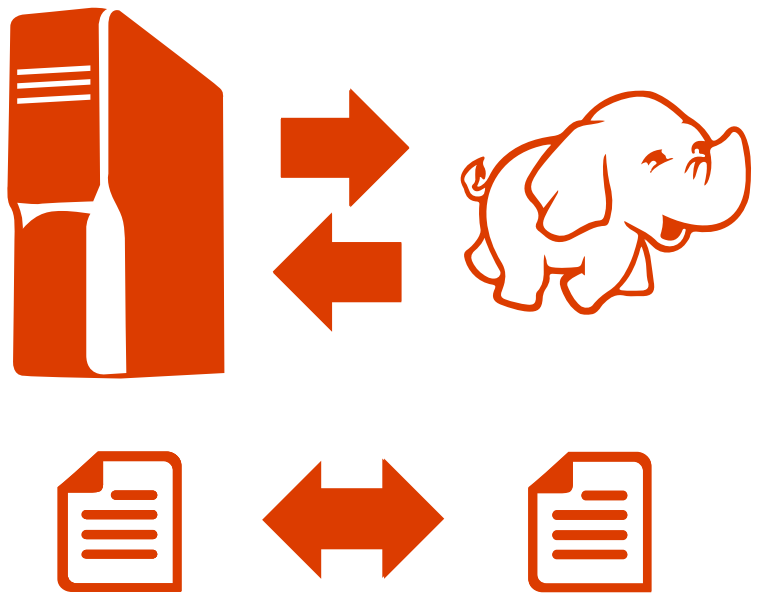


RDBMS 3.6x cheaper

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

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

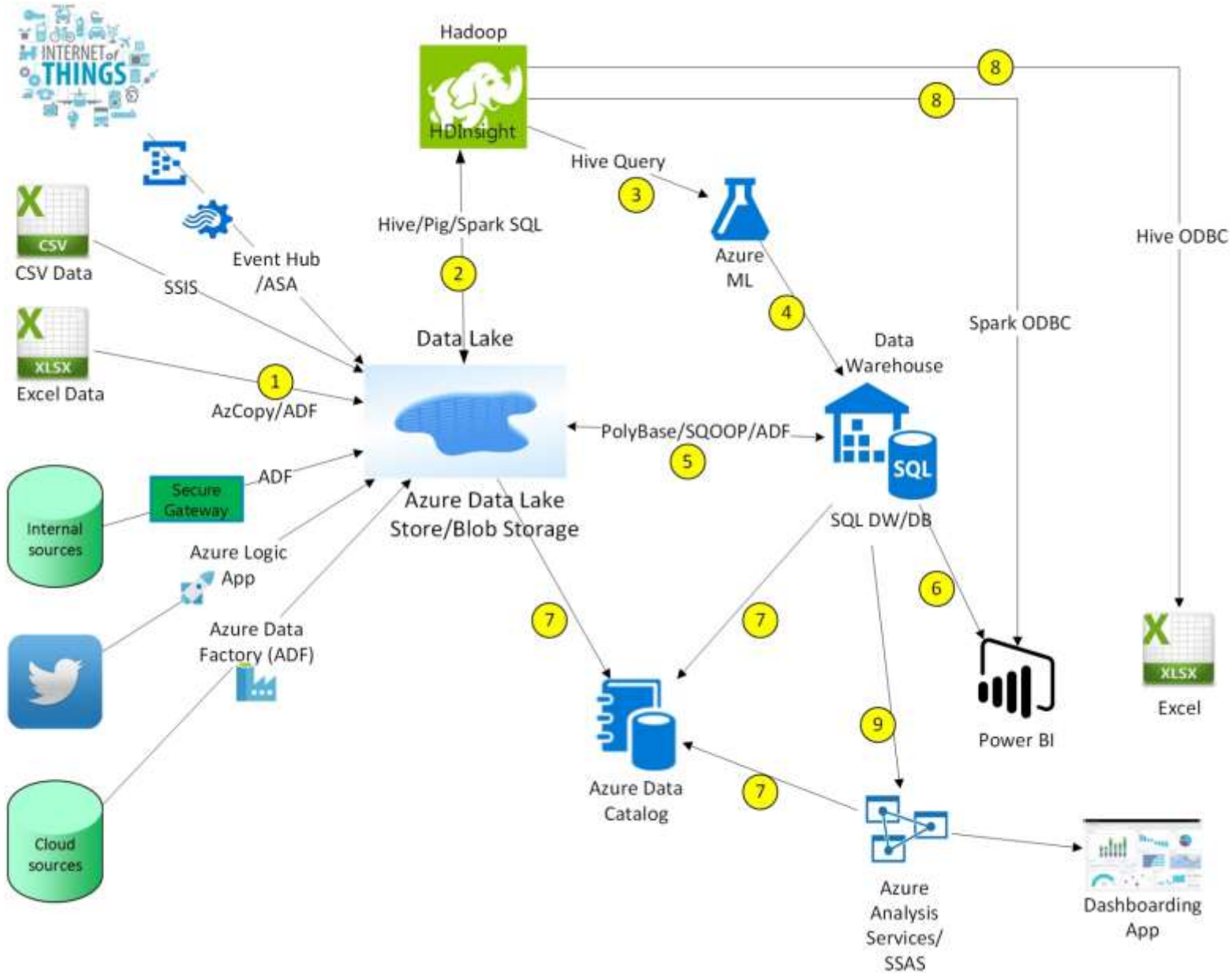
Think about future needs:

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

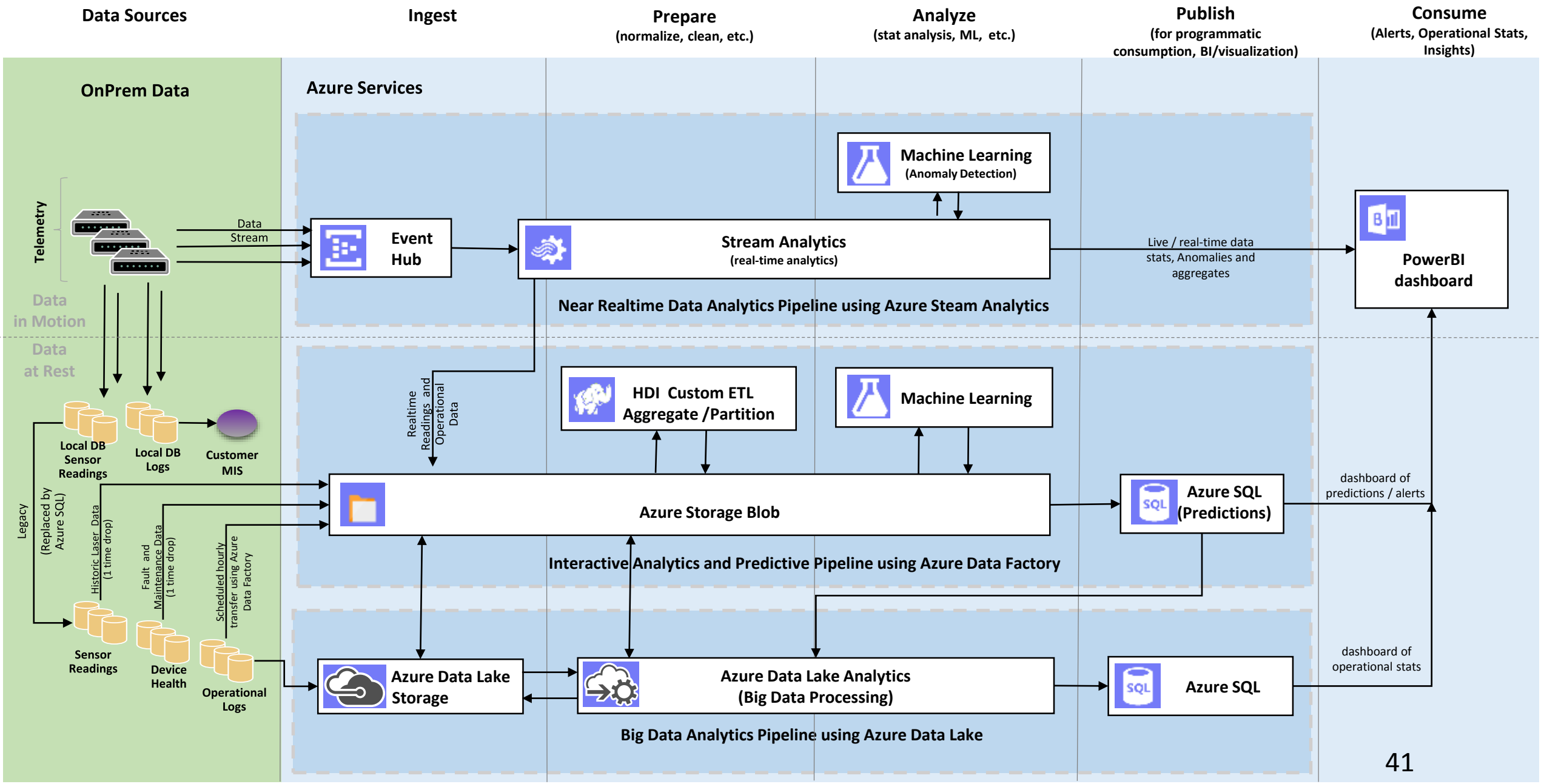
The Dream

Modern Data Warehouse

The Reality


INTERNET of THINGS
Hadoop
HDInsight
Hive Query
Hive/Pig/Spark SQL
Azure ML
Hive ODBC
Spark ODBC
CSV
CSV Data
SSIS
Event Hub /ASA
XLSX
Excel Data
AzCopy/ADF
Data Lake
Data Warehouse
PolyBase/SQOOP/ADF
SQL
SQL DW/DB
Azure Data Lake Store/Blob Storage
ADF
Secure Gateway
Internal sources
Azure Logic App
Azure Data Factory (ADF)
Cloud sources
Azure Data Catalog
Power BI
Excel
Azure Analysis Services/ SSAS
Dashboarding App
1
2
3
4
5
6
7
8
9

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)

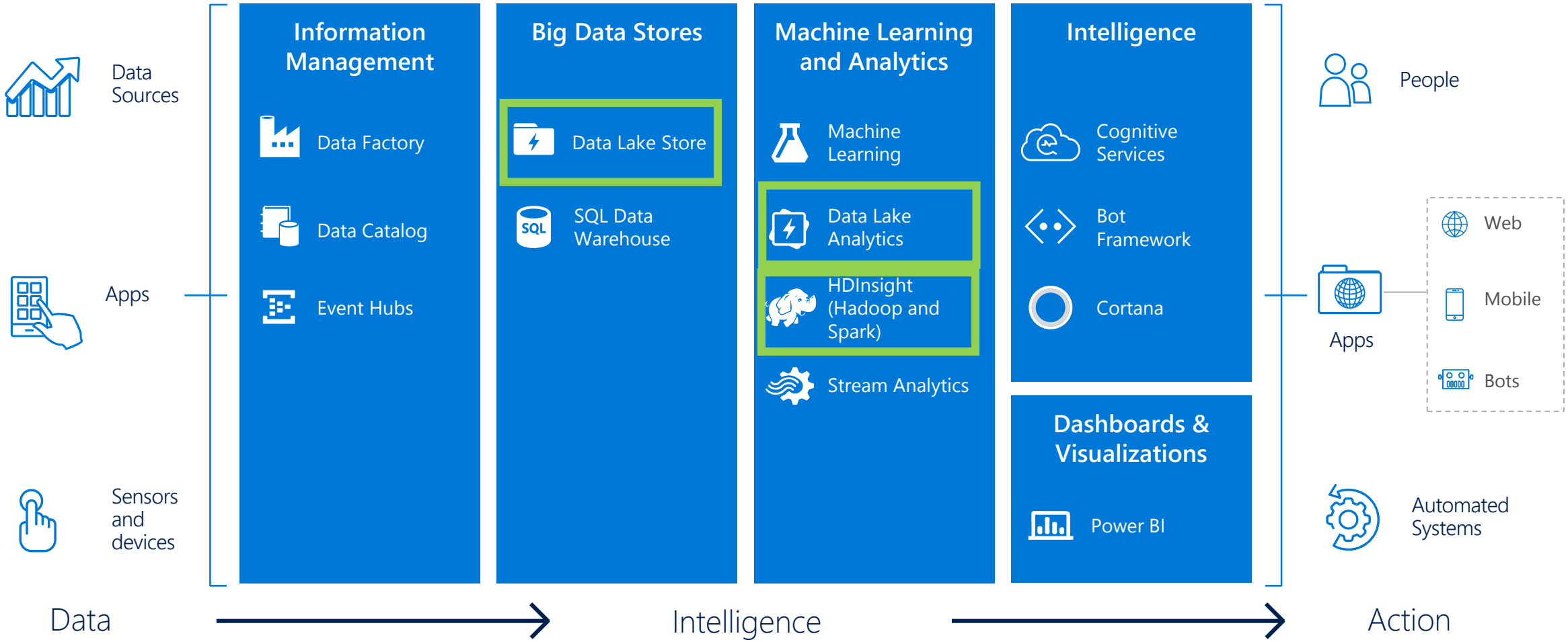
- Low latency
- High number of users
- Additional security
- Large support for tools
- Easily create reports (Self-service BI)
- *A data lake is just a glorified file folder with data files in it – how many end-users can accurately create reports from it?*

Microsoft data platform solutions

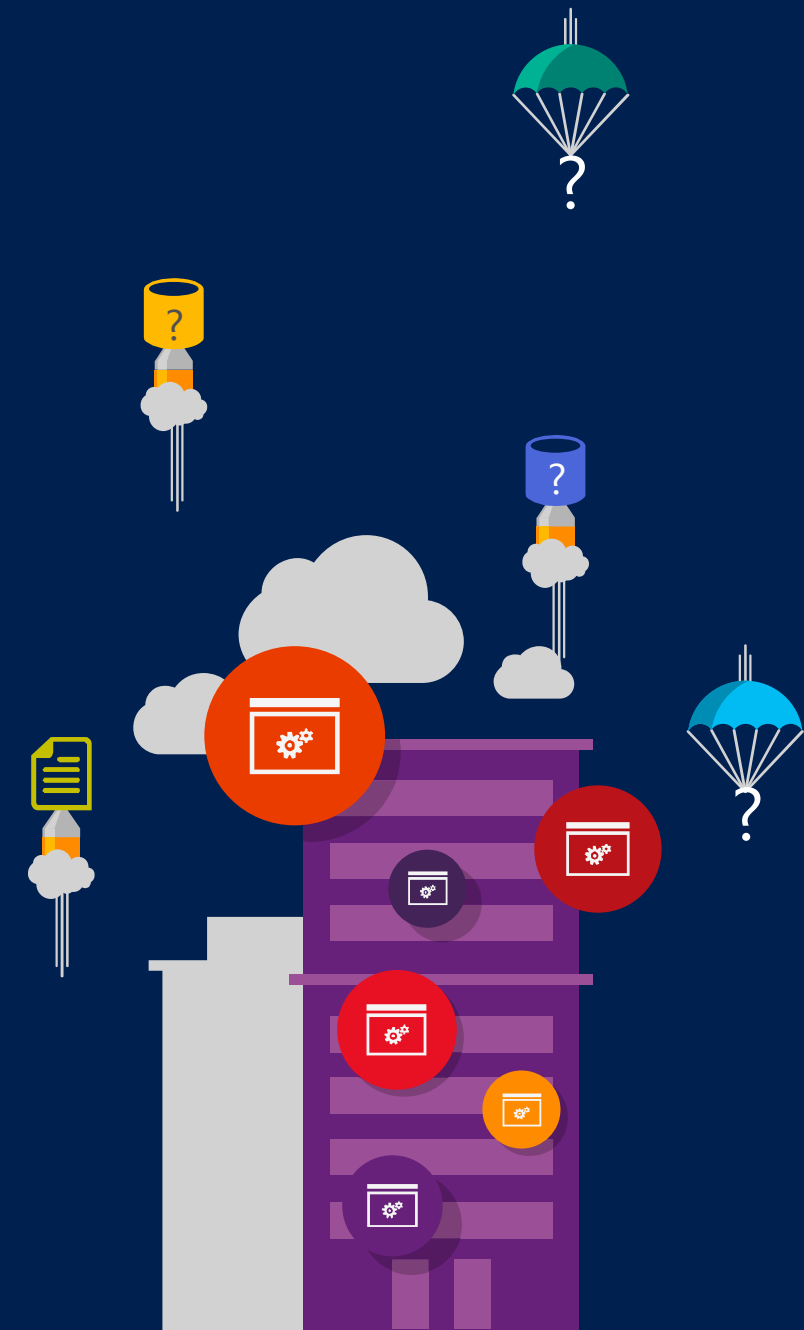
Product	Category	Description	More Info
SQL Server 2016	RDBMS	Earned top spot in Gartner's Operational Database magic quadrant. ***JSON support***	https://www.microsoft.com/en-us/server-cloud/products/sql-server-2016/
SQL Database	RDBMS/DBaaS	Cloud-based service that is provisioned and scaled quickly. Has built-in high availability and disaster recovery. ***JSON support***	https://azure.microsoft.com/en-us/services/sql-database/
SQL Data Warehouse	MPP RDBMS/DBaaS	Cloud-based service that handles relational big data. Provision and scale quickly. Can pause service to reduce cost	https://azure.microsoft.com/en-us/services/sql-data-warehouse/
Analytics Platform System (APS)	MPP RDBMS	Big data analytics appliance for high performance and seamless integration of all your data	https://www.microsoft.com/en-us/server-cloud/products/analytics-platform-system/
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	https://azure.microsoft.com/en-us/services/data-lake-store/
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	https://azure.microsoft.com/en-us/services/data-lake-analytics/
HDInsight	PaaS Hadoop compute	A managed Apache Hadoop, Spark, R, HBase, and Storm cloud service made easy	https://azure.microsoft.com/en-us/services/hdinsight/
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	https://azure.microsoft.com/en-us/services/documentdb/
Azure Table Storage	PaaS NoSQL: Key-value Store	Store large amount of semi-structured data in the cloud	https://azure.microsoft.com/en-us/services/storage/tables/

Cortana Intelligence Suite

Integrated as part of an end-to-end suite



Federated Querying



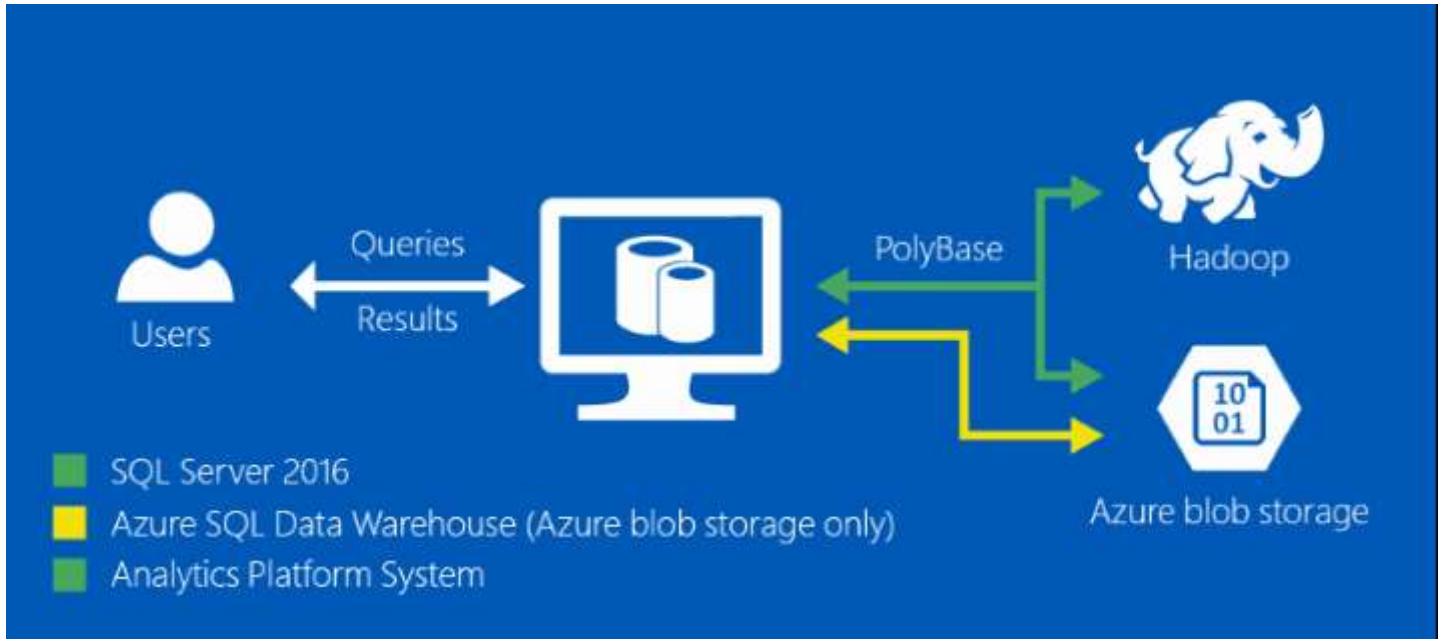
Federated Querying

Other names: Data virtualization, logical data warehouse, data federation, virtual database, and decentralized data warehouse.

A model that allows a single query to retrieve and combine data as it sits from multiple data sources, so as to not need to use ETL or learn more than one retrieval technology

PolyBase

Query relational and non-relational data with T-SQL



By preview early this year PolyBase will add support for Teradata, Oracle, SQL Server, MongoDB, and generic ODBC (Spark, Hive, Impala, DB2)

Vs U-SQL: PolyBase is interactive while U-SQL is batch. U-SQL more code to query data but more formats (JSON) and libraries/UDOs and supports writes to blob/ADLS

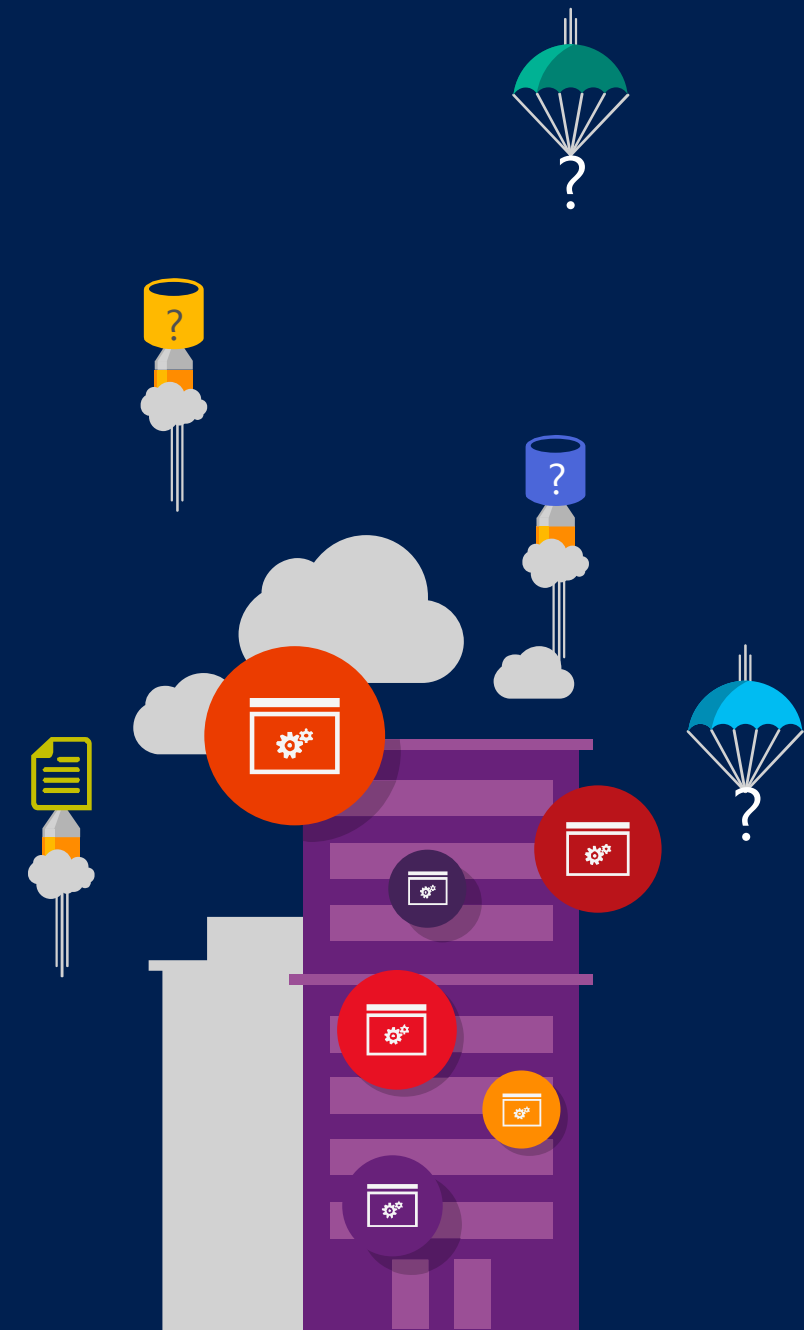
Capability

T-SQL for querying relational and non-relational data across SQL Server (APS, SQL Server 2016, SQL DW) and Hadoop and Azure blob storage (soon ADLS)

Benefits

- New business insights across your data lake
- Leverage existing skillsets and BI tools
- Faster time to insights and simplified ETL process

Solution in the cloud



Benefits of the cloud

Agility

- Unlimited elastic scale
- Pay for what you need

Innovation

- Quick "Time to market"
- Fail fast

Risk

- Availability
- Reliability
- Security

Total cost of ownership calculator: https://www.tco.microsoft.com/

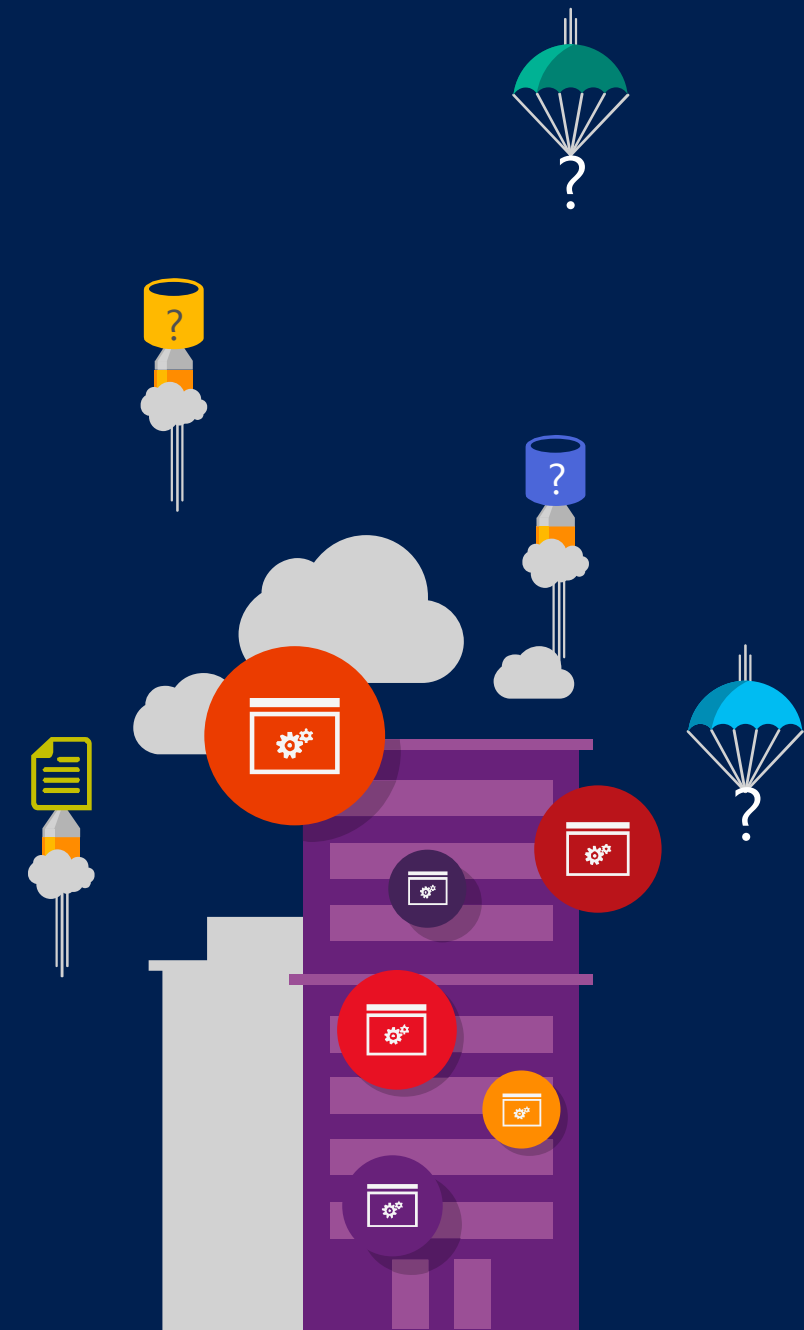
Constraints of on-premise data

- Scale constrained to on-premise procurement
- Capex up-front costs, most companies instead prefer a yearly operating expense (OpEx)
- A staff of employees or consultants must be retained to administer and support the hardware and software in place
- Expertise needed for tuning and deployment

Talking points when using the cloud for DW

- Public and private cloud
- Cloud-born data vs on-prem born data
- Transfer cost from/to cloud and on-prem
- Sensitive data on-prem, non-sensitive in cloud
- Look at hybrid solutions

SMP vs MPP



SMP vs MPP

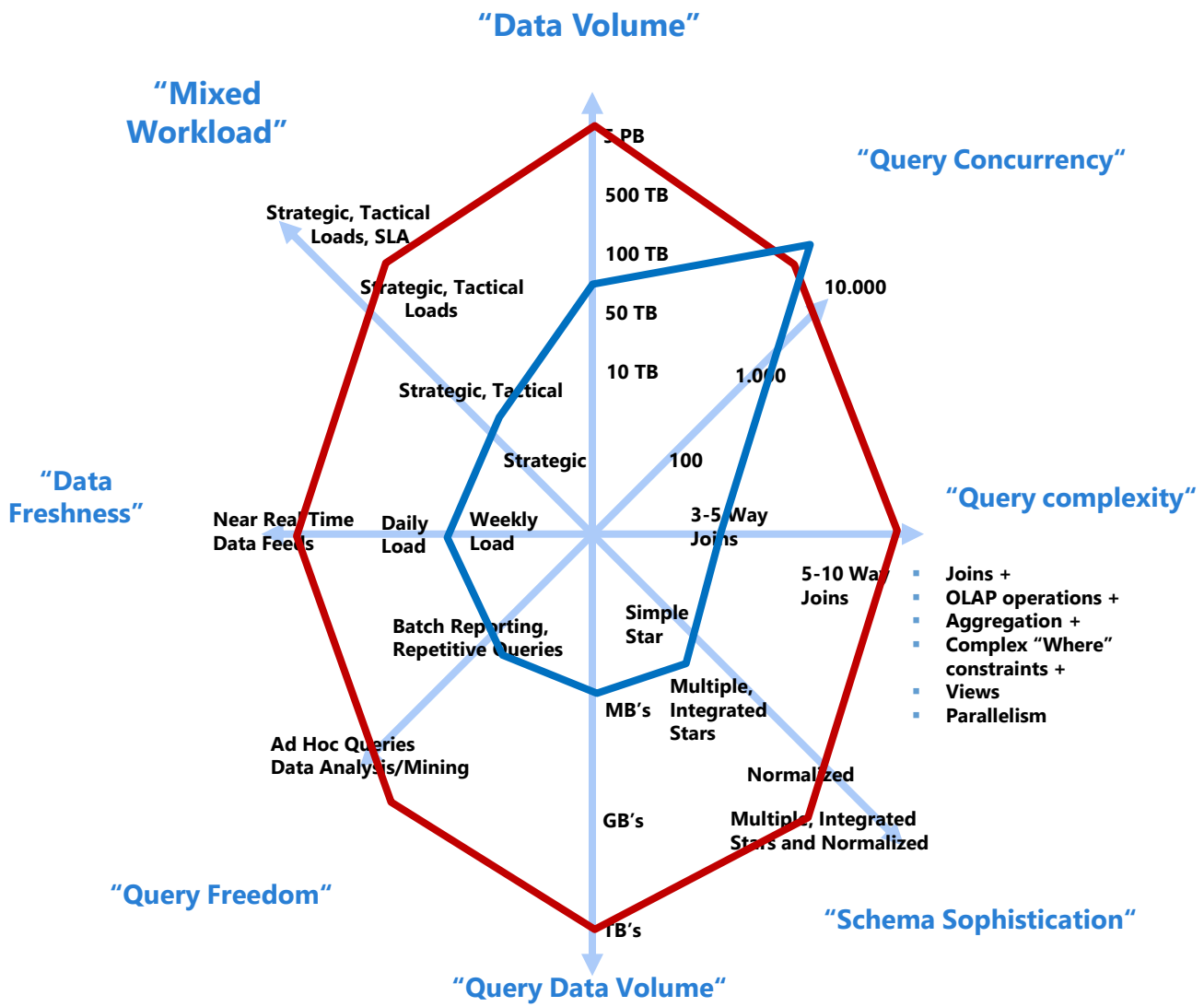
SMP - Symmetric Multiprocessing

- Multiple CPUs used to complete individual processes simultaneously
- All CPUs share the same memory, disks, and network controllers (scale-up)
- All SQL Server implementations up until now have been SMP
- Mostly, the solution is housed on a shared SAN

MPP - Massively Parallel Processing

- Uses many separate CPUs running in parallel to execute a single program
- Shared Nothing: Each CPU has its own memory and disk (scale-out)
- Segments communicate using high-speed network between nodes

DW SCALABILITY SPIDER CHART



MPP – Multidimensional Scalability

SMP – Tunable in one dimension on cost of other dimensions

The spiderweb depicts important attributes to consider when evaluating Data Warehousing options.

Big Data support is newest dimension.

Summary

- We live in an increasingly data-intensive world
- Much of the data stored online and analyzed today is more varied than the data stored in recent years
- More of our data arrives in near-real time
- “Data is the new currency!”

This present a large business opportunity. Are you ready for it?

Other Related Presentations

- Building a Big Data Solution
- Choosing technologies for a big data solution in the cloud
- How does Microsoft solve Big Data?
- Benefits of the Azure cloud
- Should I move my database to the cloud?
- Implement SQL Server on a Azure VM
- Relational databases vs Non-relational databases
- Introduction to Microsoft's Hadoop solution (HDInsight)
- Introducing Azure SQL Database
- Introducing Azure SQL Data Warehouse

Visit my blog at: JamesSerra.com (where these slide decks are posted under the "Presentation" tab)

Resources

- Why use a data lake? http://bit.ly/1WDy848
- Big Data Architectures http://bit.ly/1RBbAbS
- The Modern Data Warehouse: http://bit.ly/1xuX4Py
- Hadoop and Data Warehouses: http://bit.ly/1xuXfu9

Q & A

James Serra, Big Data Evangelist
Email me at: JamesSerra3@gmail.com
Follow me at: @JamesSerra
Link to me at: www.linkedin.com/in/JamesSerra
Visit my blog at: JamesSerra.com (where this slide deck is posted under the “Presentations” tab)