Hitachi Consulting
We make it happen. Better.

Enterprise Cloud Data Platforms

with Microsoft Azure & Cortana Analytical Suite

Khalid M. Salama, Ph.D.
Business Insights & Analytics
Hitachi Consulting UK

 |

Outline

Hitachi Consulting
We make it happen. Better.

- Microsoft Azure and Cloud Computing
- Azure Data Architecture and Services
- EDW with Azure SQL Data Warehouse
- Big Data Storage and Analytics on Azure Data Lake and HDInsight
- Data Processing and Movements with Azure Data Factory
- Real-time Processing with Azure Event Hubs and Stream Analytics
- Data Science and Machine Learning on Microsoft Azure
- Data Discoverability with Azure Data Catalog

Hitachi Consulting
We make it happen. Better.

Microsoft Azure and Cloud Computing

 |

Hitachi Consulting
We make it happen. Better.

Cloud Computing

Enabling the future of computing

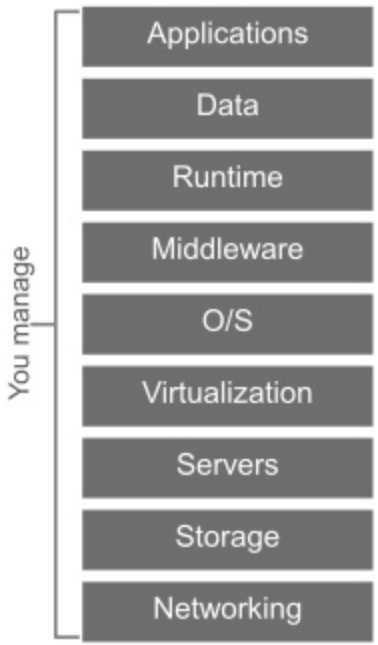
On-demand Self-service	• Web interfaces that deliver capacity in seconds or minutes
Broad Network Access	• Support many devices, such as mobile phones, laptops, tablets and workstations
Resource Pooling	• Thousands to hundred of thousands severs are immediately usable
Rapid Elasticity	• On-demand or automatic (threshold-based) scale-up or down.
Measured Service	• Pay as you use, what you use, Improve operational metric and cost accounting
Innovation	• Modern IT capabilities and services provided frequently

Cloud Computing on Microsoft Azure

Computing Models

Hitachi Consulting
We make it happen. Better.

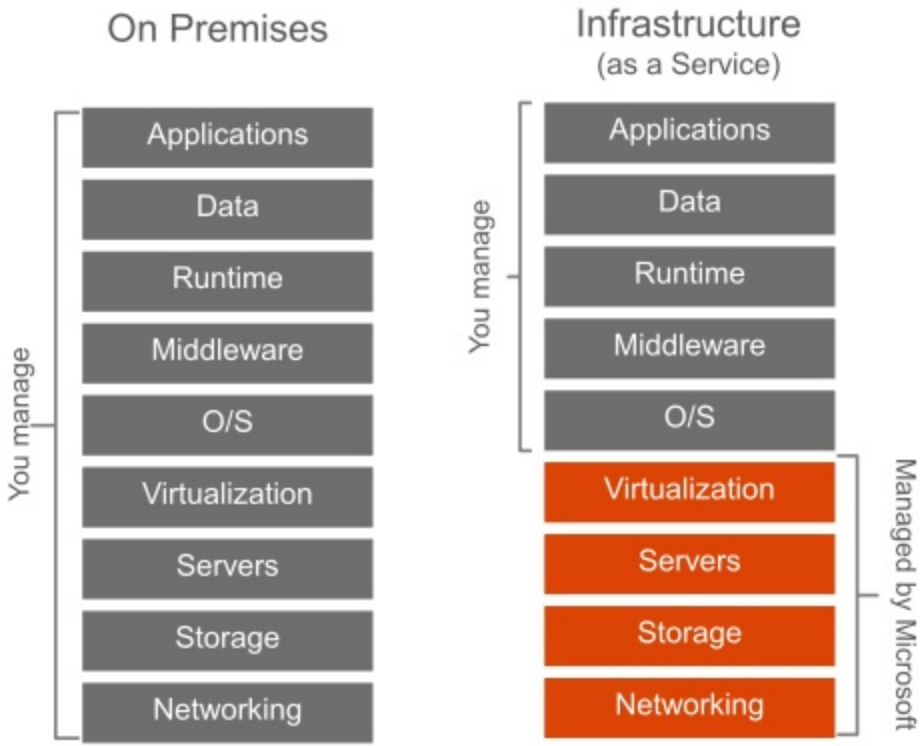
On Premises



 |

Cloud Computing on Microsoft Azure

Computing Models

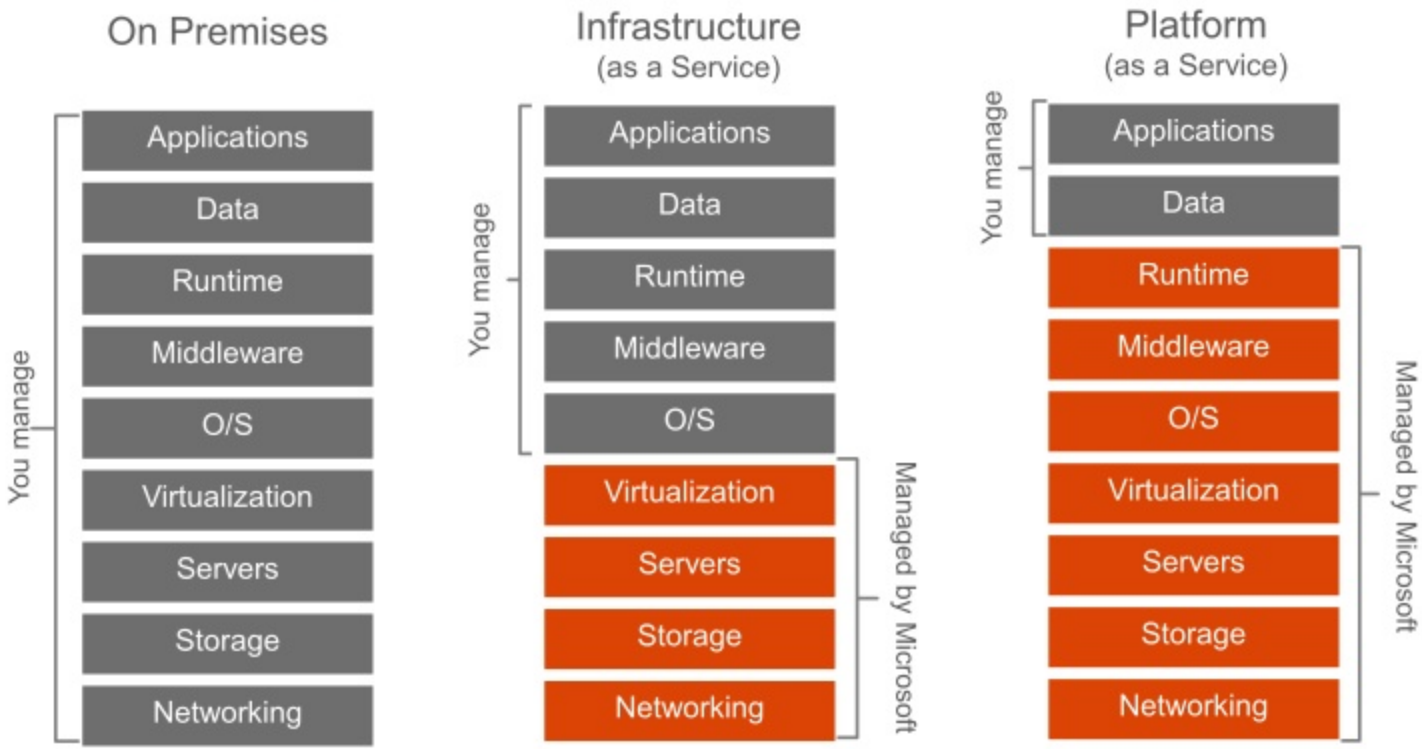


 |

Hitachi Consulting
We make it happen. Better.

Cloud Computing on Microsoft Azure

Computing Models

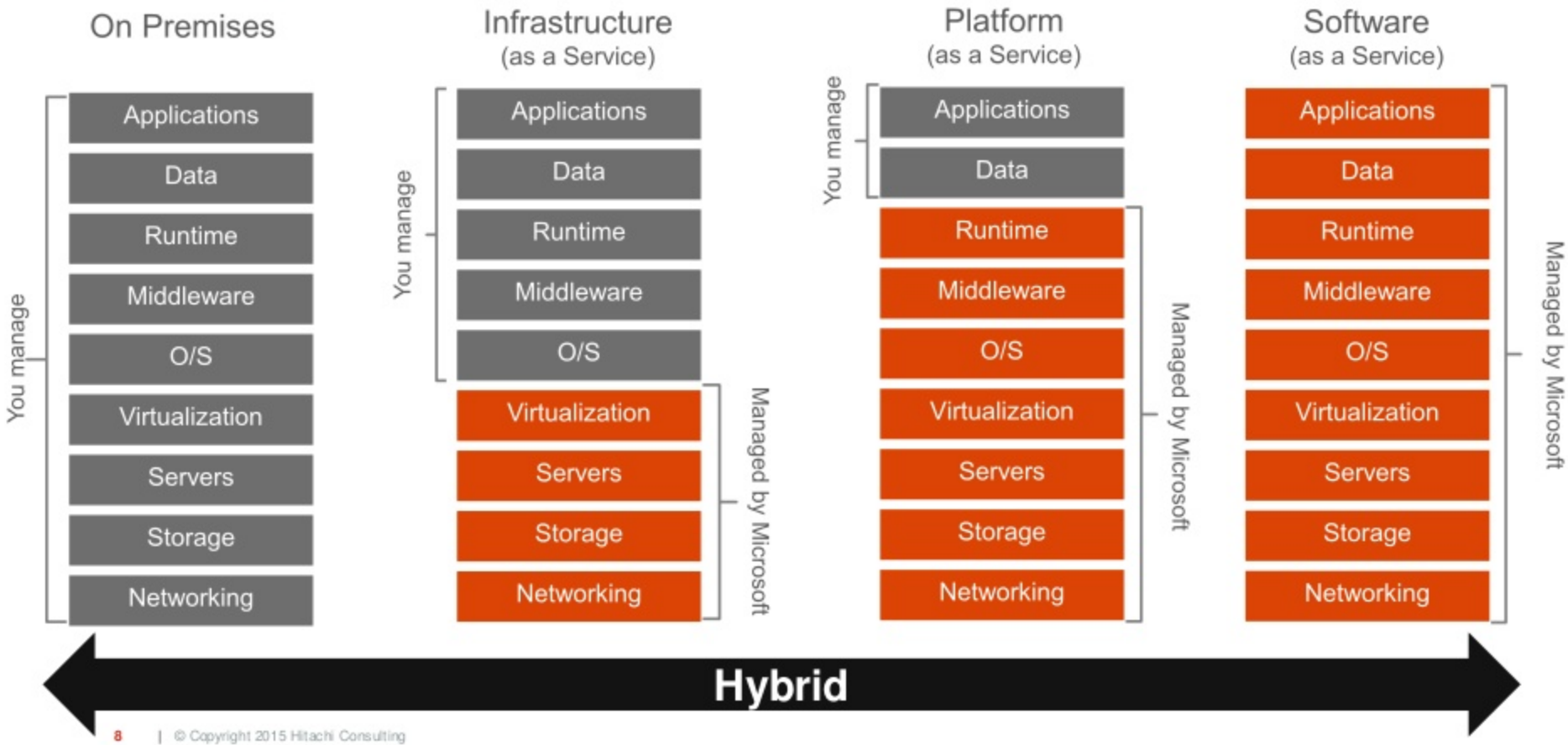


 |

Cloud Computing on Microsoft Azure

Computing Models

Hitachi Consulting
We make it happen. Better.



 |

Cloud Computing on Microsoft Azure

Computing Models - Examples

Hitachi Consulting
We make it happen. Better.

- **Infrastructure as a Service (IaaS)** – VMs and Networking Services
  (build Software infrastructure, develop apps, use apps, Software maintained infrastructure)

- **Platform as a Service (PaaS)** – Data and App Services
  (develop apps, use apps)

- **Software as a Service (SaaS)** – SharePoint Online, Office 365, Microsoft Dynamics, Visual Studio Online, Azure Cognitive Services, etc.
  (use apps)

Cloud Computing on Microsoft Azure

So why the cloud?

Hitachi Consulting
We make it happen. Better.

Agility

- Configure and provision in minutes
- Automatic/on-demand scaling up/down

Economics

- Pay as you use (100% Utilization)
- Elastic capacity

Innovation

- Connectivity
- Innovative capabilities and services
- New compute models

Hitachi Consulting
We make it happen. Better.

Azure Data Architecture and Services

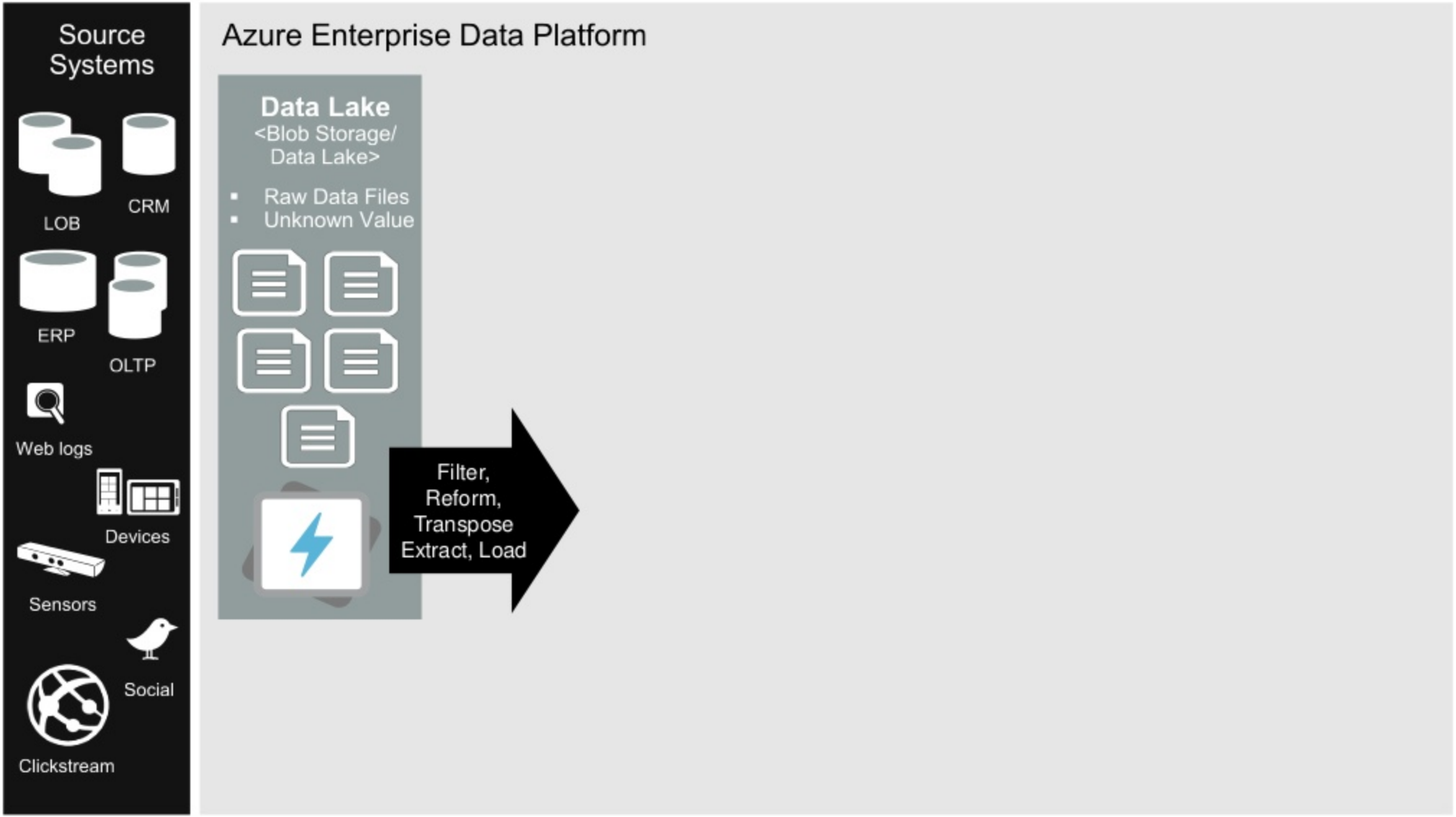
Hitachi Consulting
We make it happen. Better.

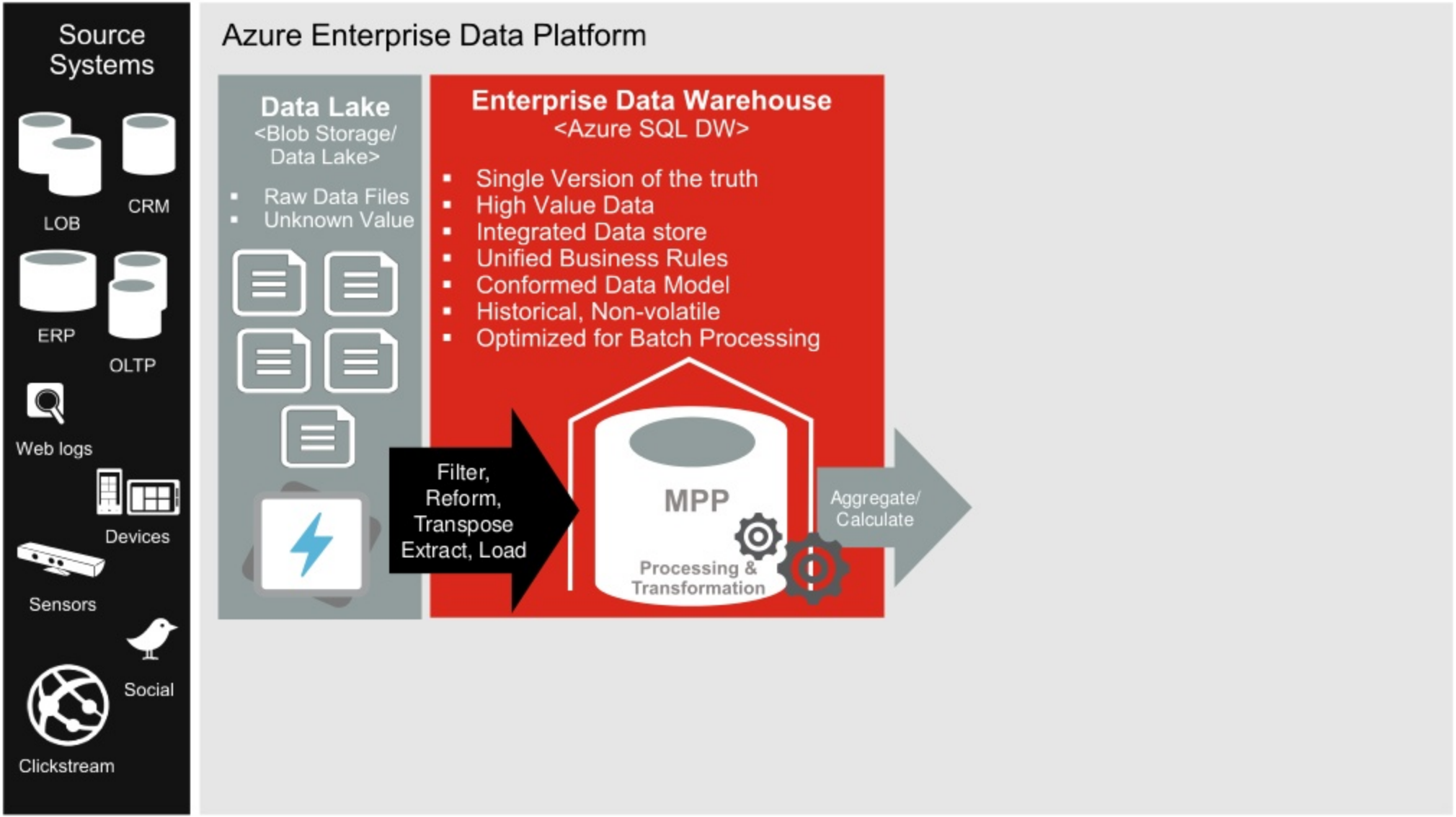
Enterprise Data Platforms

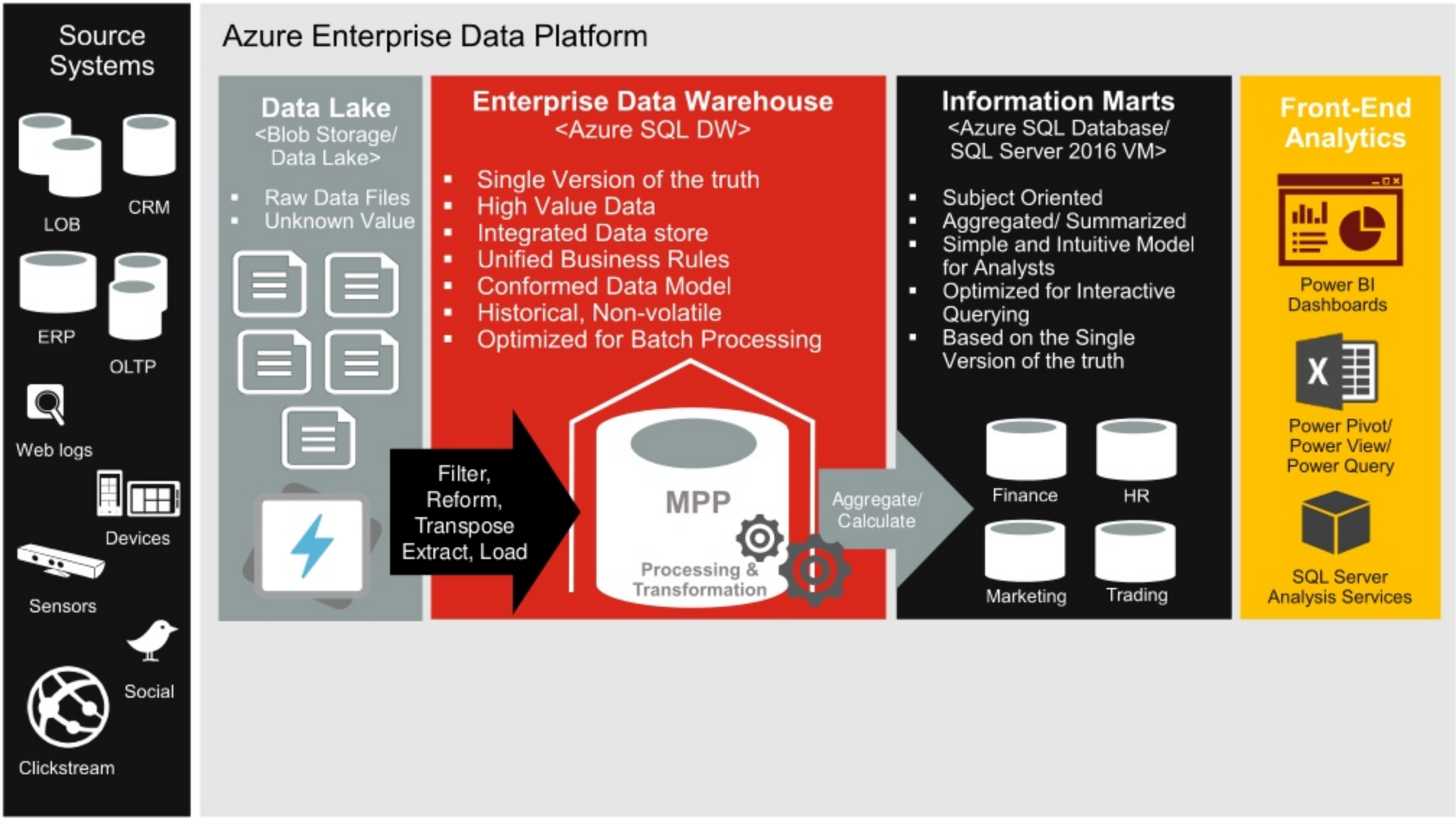
Challenges of Enterprise Information Management

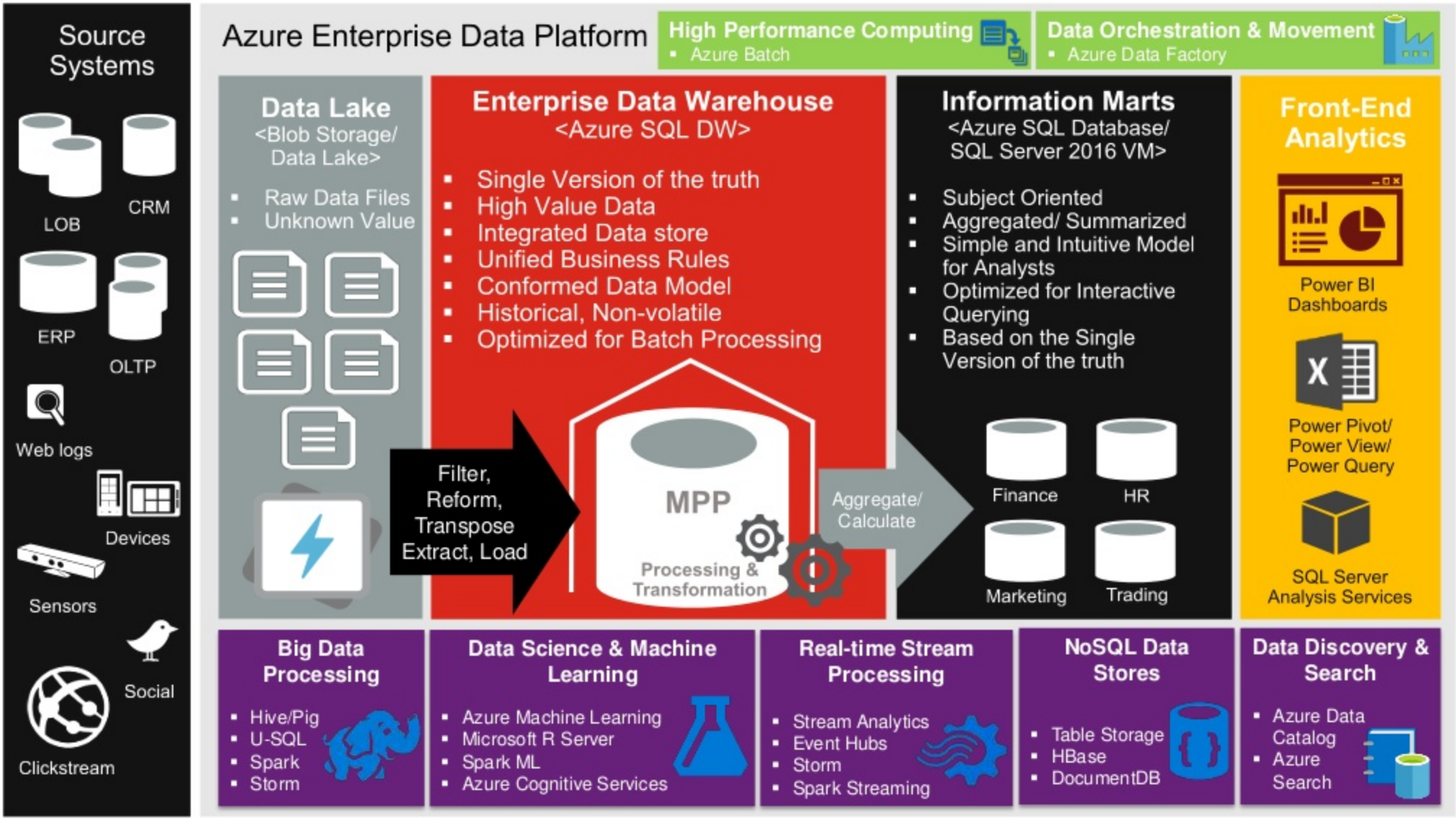
 |


Source Systems
LOB
CRM
ERP
OLTP
Web logs
Devices
Sensors
Social
Clickstream
Azure Enterprise Data Platform


Source Systems
LOB
CRM
ERP
OLTP
Web logs
Devices
Sensors
Social
Clickstream
Azure Enterprise Data Platform
Data Lake
<Blob Storage/ Data Lake>
Raw Data Files
Unknown Value
Filter, Reform, Transpose Extract, Load


Source Systems
LOB
CRM
ERP
OLTP
Web logs
Devices
Sensors
Social
Clickstream
Azure Enterprise Data Platform
Data Lake
<Blob Storage/ Data Lake>
Raw Data Files
Unknown Value
Filter, Reform, Transpose Extract, Load
Enterprise Data Warehouse
<Azure SQL DW>
Single Version of the truth
High Value Data
Integrated Data store
Unified Business Rules
Conformed Data Model
Historical, Non-volatile
Optimized for Batch Processing
MPP
Processing & Transformation
Aggregate/ Calculate

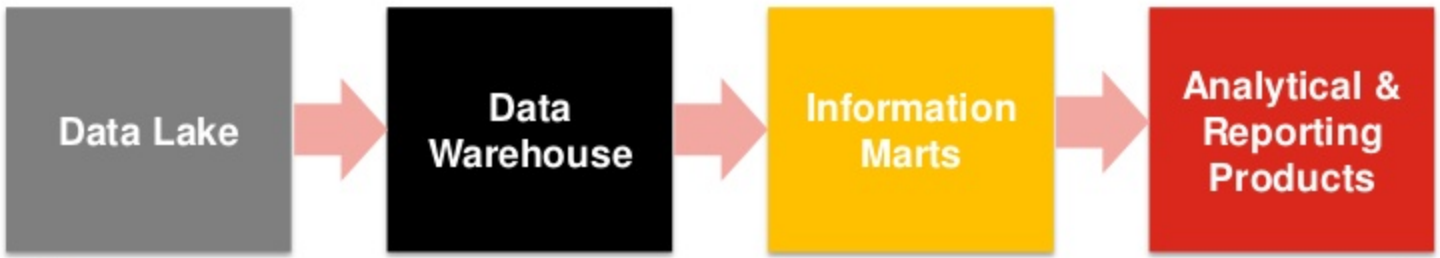

Source Systems
LOB
CRM
ERP
OLTP
Web logs
Devices
Sensors
Social
Clickstream
Azure Enterprise Data Platform
Data Lake
<Blob Storage/ Data Lake>
Raw Data Files
Unknown Value
Filter, Reform, Transpose Extract, Load
Enterprise Data Warehouse
<Azure SQL DW>
Single Version of the truth
High Value Data
Integrated Data store
Unified Business Rules
Conformed Data Model
Historical, Non-volatile
Optimized for Batch Processing
MPP
Processing & Transformation
Aggregate/ Calculate
Information Marts
<Azure SQL Database/ SQL Server 2016 VM>
Subject Oriented
Aggregated/ Summarized
Simple and Intuitive Model for Analysts
Optimized for Interactive Querying
Based on the Single Version of the truth
Finance
HR
Marketing
Trading
Front-End Analytics
Power BI Dashboards
Power Pivot/ Power View/ Power Query
SQL Server Analysis Services


Source Systems
LOB
CRM
ERP
OLTP
Web logs
Devices
Sensors
Social
Clickstream
Azure Enterprise Data Platform
High Performance Computing
Azure Batch
Data Orchestration & Movement
Azure Data Factory
Data Lake
<Blob Storage/ Data Lake>
Raw Data Files
Unknown Value
Filter, Reform, Transpose Extract, Load
Enterprise Data Warehouse
<Azure SQL DW>
Single Version of the truth
High Value Data
Integrated Data store
Unified Business Rules
Conformed Data Model
Historical, Non-volatile
Optimized for Batch Processing
MPP
Processing & Transformation
Aggregate/ Calculate
Information Marts
<Azure SQL Database/ SQL Server 2016 VM>
Subject Oriented
Aggregated/ Summarized
Simple and Intuitive Model for Analysts
Optimized for Interactive Querying
Based on the Single Version of the truth
Finance
HR
Marketing
Trading
Front-End Analytics
Power BI Dashboards
Power Pivot/ Power View/ Power Query
SQL Server Analysis Services
Big Data Processing
Hive/Pig
U-SQL
Spark
Storm
Data Science & Machine Learning
Azure Machine Learning
Microsoft R Server
Spark ML
Azure Cognitive Services
Real-time Stream Processing
Stream Analytics
Event Hubs
Storm
Spark Streaming
NoSQL Data Stores
Table Storage
HBase
DocumentDB
Data Discovery & Search
Azure Data Catalog
Azure Search

Azure Enterprise Data Platform

Layers and relevance

Hitachi Consulting
We make it happen. Better.

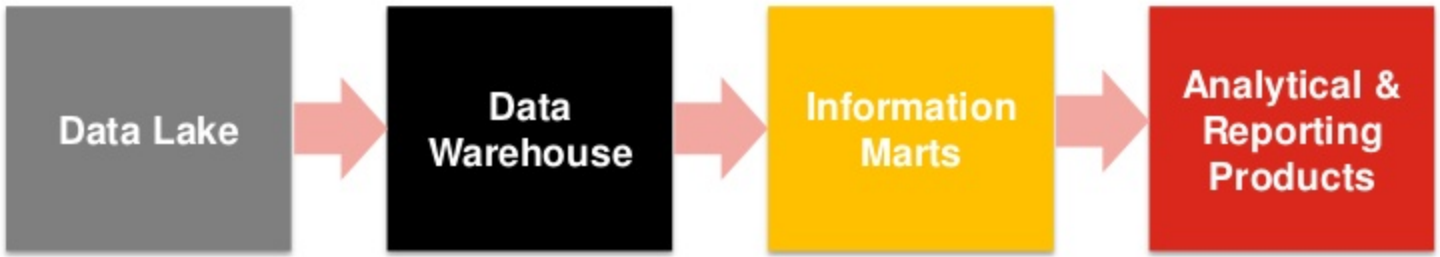


Raw - Unstructured	Data	Processed - Structured
Low - Unknown	Value	High-value
Data → Structure	Approach	Structure → Data
Batch	Query	Interactive
Strict /Limited	Access	Available/ Wide
Big (Everything)	Size	Small (Focused)
Data Professionals	Users	Information Workers

 |

Azure Enterprise Data Platform

Business Agility & IT Governance

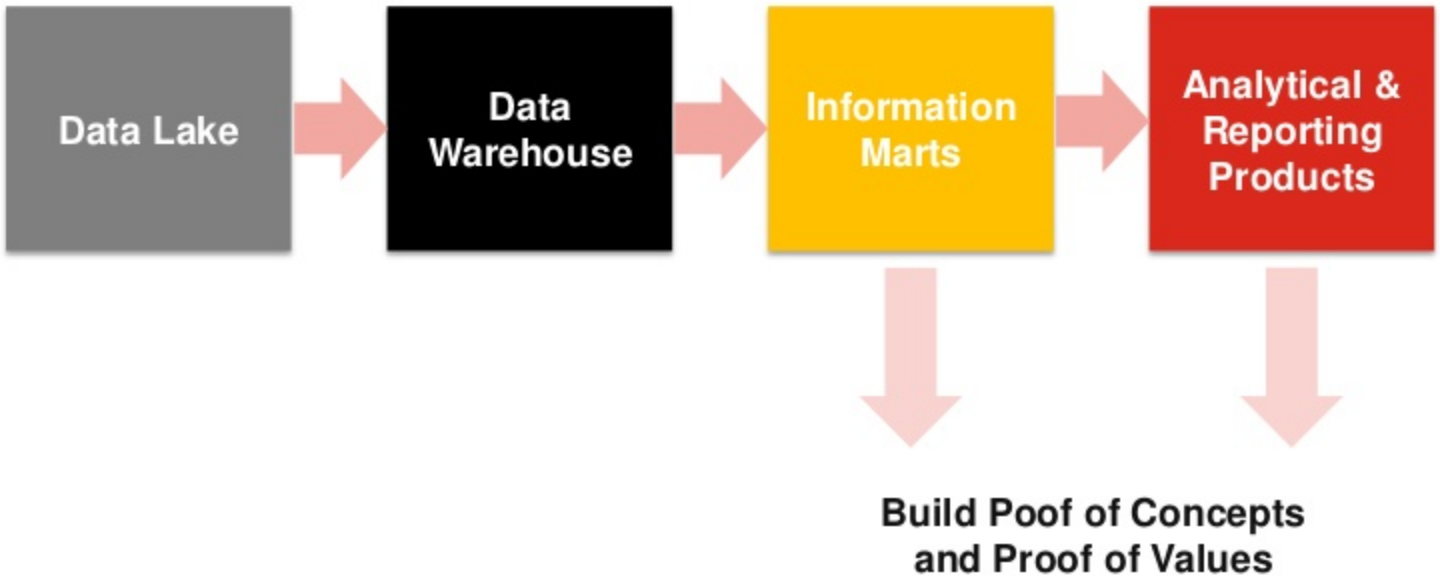


 |

Hitachi Consulting
We make it happen. Better.

Azure Enterprise Data Platform

Business Agility & IT Governance

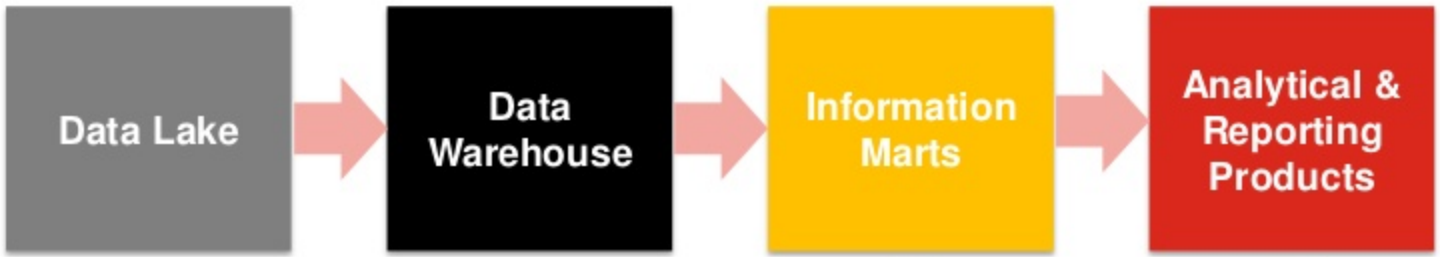


 |

Azure Enterprise Data Platform

Business Agility & IT Governance

Hitachi Consulting
We make it happen. Better.



Build Poof of Concepts and Proof of Values

- Build a show-case Power Pivot Model
- Spin up an experimental VM or Azure Services
- Import new data to a sandbox area in the mart
- Delve into data lake for exploratory data analysis

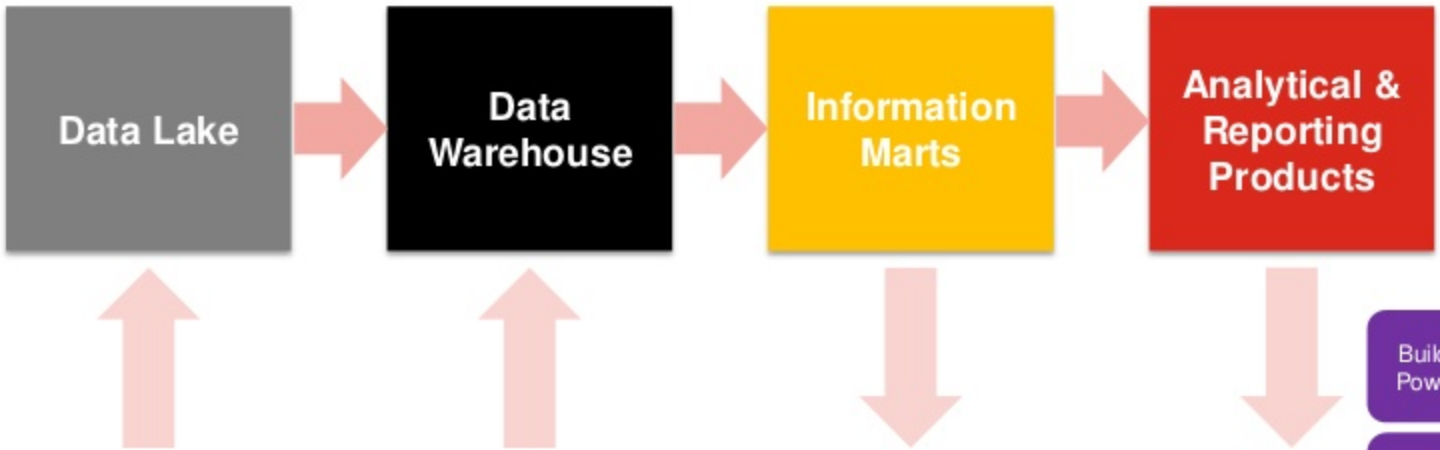
Data Analysts and Data Scientists

 |

Azure Enterprise Data Platform

Business Agility & IT Governance

Hitachi Consulting
We make it happen. Better.



 |

Hitachi Consulting
We make it happen. Better.

Azure Data Lake Storage and Analytics

Hitachi Consulting
We make it happen. Better.

Data Lake – The Concept

What is that? And Why it is needed?

What is a Data lake

A repository that holds raw data file extracts of all the Enterprise source systems

Usually a Distributed File System (HDFS, Blob Storage, Amazon S3, GFS, etc.) to support Big Data Processing

Pentaho (Hitachi Data Systems) co-founder and CTO, James Dixon coined the term 'Data Lake'

Why a Data Lake

Single and cheap repository of Enterprise-wide (raw) data

Abstracts your DW/ Business Intelligence Solutions from source systems

Supports potential Big Data Processing and Analytics

Data Lake – The Concept

Data Processing

Hitachi Consulting
We make it happen. Better.

Data stored in its native form (Structured, Semi-structured, and Unstructured):
Text Files, JSON, XML, Logs, Mainframe, etc.

Big Data technologies, such as MapReduce, Pig, Hive, Spark, and U-SQL, are used to process the data in a procedural fashion:

Transpose/ reshape data into a tabular form	**E**xtract/filter valuable and relevant data elements	**L**oad prepared datasets into a relational EDW

 |

Hitachi Consulting
We make it happen. Better.

Azure Data Lake

Azure Data Lake Services

HDInsight

- Hadoop Cluster as a Service
- MapReduce | Pig | Hive | Spark HBase | Oozie | Storm
- Unit of Cost: Running nodes

Data Lake Analytics

- Big Data Jobs as a Service
- U-SQL
- Unit of Cost: running Job

Data Lake Storage

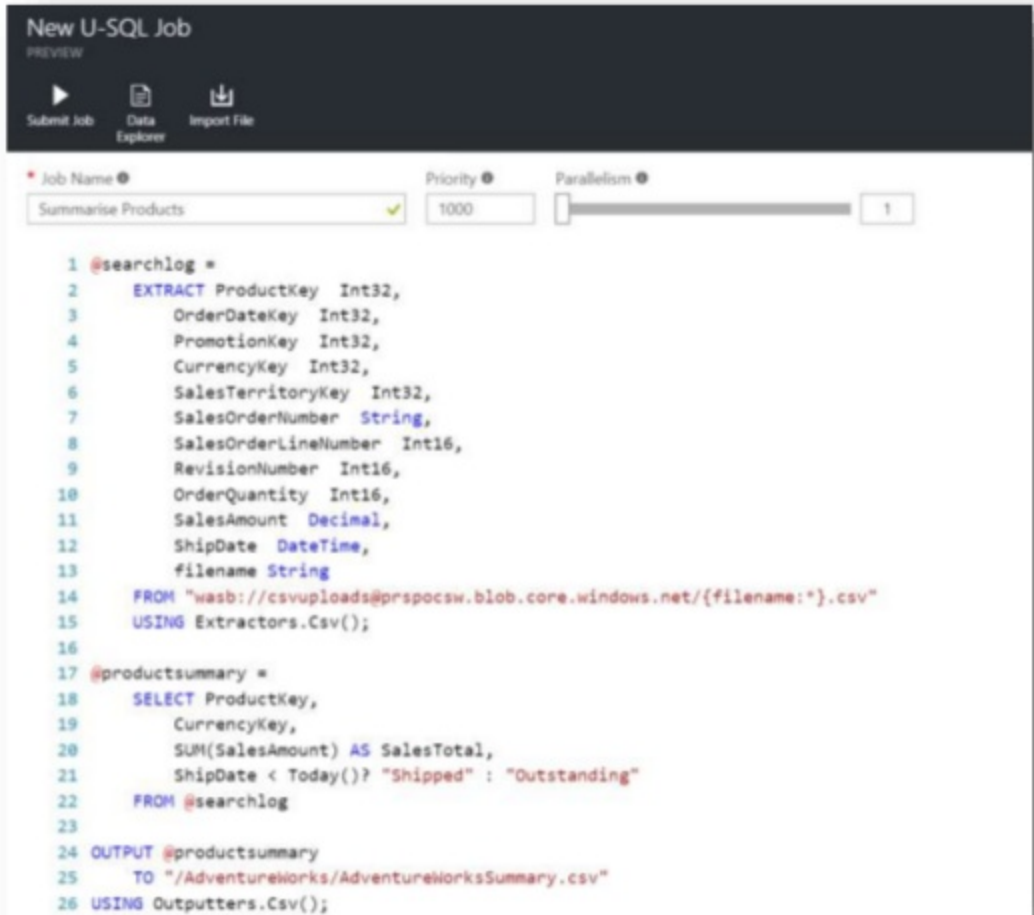
- Built for data files, rather than Blobs
- WebHDFS REST APIs for Hadoop integration
- Optimized for analytical workloads
- Unlimited Storage, petabyte files

Azure Data Lake Analytics

U-SQL

Hitachi Consulting
We make it happen. Better.

- C# meets SQL (.NET data type, C# expressions + SQL set statements)
- Natively Parallel
- Set-based and procedural processing
- Works with structured and unstructured data
- Extensible (User-defined Functions, Operators and Aggregate Functions)
- Reuse of custom build .NET assemblies

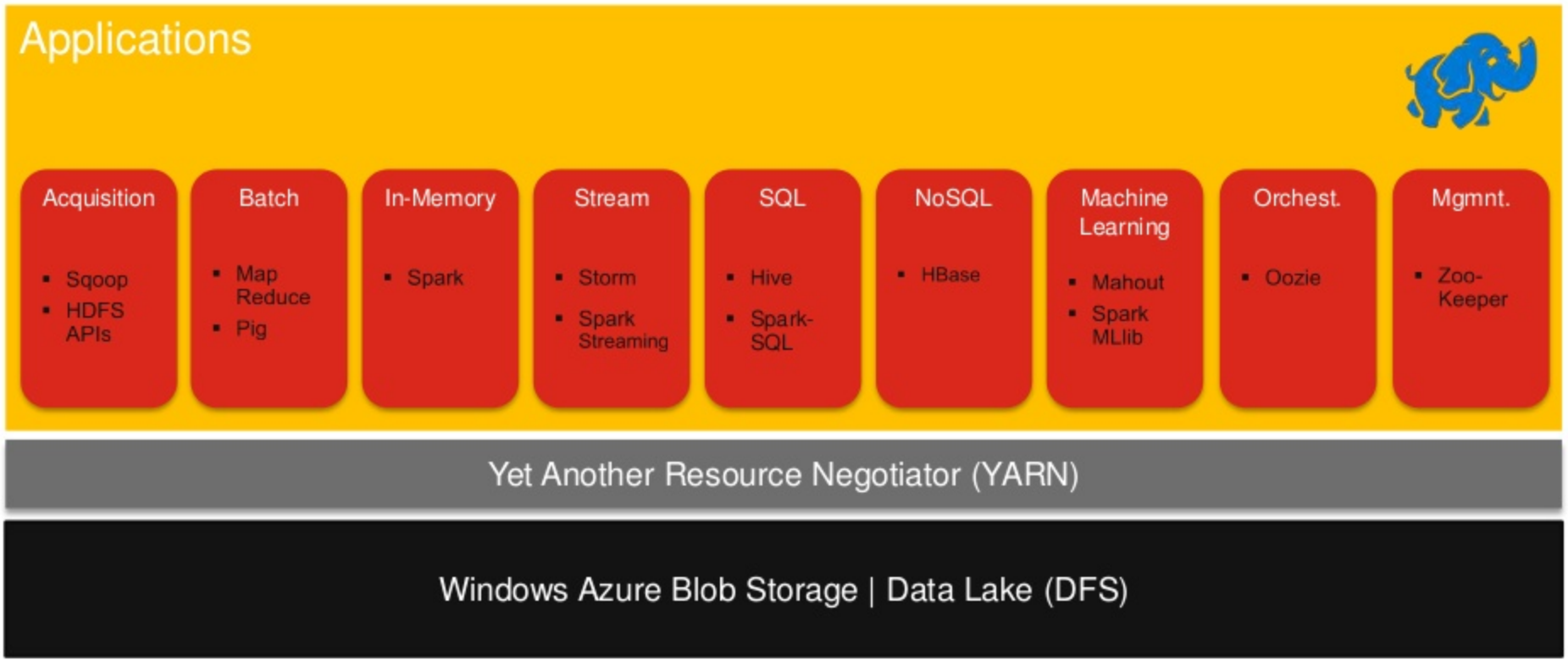


 |

Hitachi Consulting
We make it happen. Better.

Azure HDInsight

Hadoop on Microsoft Cloud



 |

Azure SQL Data Warehouse

Azure SQL Data Warehouse

Enterprise Data Warehouse – why?

Hitachi Consulting
We make it happen. Better.

Azure SQL Data Warehouse

Enterprise Data Warehouse – why?

Hitachi Consulting
We make it happen. Better.

" That is great! Now we have a data lake that has all the data that we want. Let's spin up a VM, build a database, and produce the reports we want."

– Said Mr. quick-win from the business.

Hitachi Consulting
We make it happen. Better.

Azure SQL Data Warehouse

Enterprise Data Warehouse – why?

" That is great! Now we have a data lake that has all the data that we want. Let's spin up a VM, build a database, and produce the reports we want."

– Said Mr. quick-win from the business.

"No. The data lake has raw data with no business structure. High-value data needs to be loaded into a canonical relational form, where enterprise-wide business entities , calculations, and rules are defined. Building reporting products on top of the data lake will results in unreliable information silos without single version of the truth."

– Said Mr. enterprise data architect from the IS.

Hitachi Consulting
We make it happen. Better.

Azure SQL Data Warehouse

A Parallel DW on the cloud

An elastic, Massively Parallel Processing (MPP), cloud-based data warehouse as a service, with enterprise-class features

Suitable for

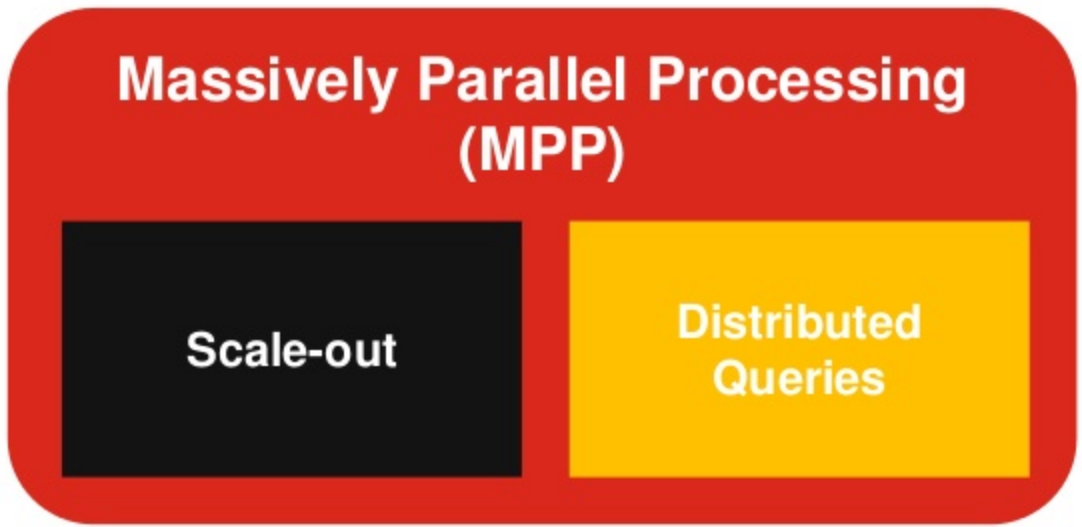
- Batch data processing workloads
- Consolidate high-value data into a single enterprise-wide structured store
- Model, transform and aggregate data
- Perform query analysis across large

Not Suitable for

- Operational workloads
- High frequency reads & writes
- Singleton operations → (Information Mart)
- Row by row processing needs → (Data Lake)
- High concurrency → (Information Mart)

Azure SQL Data Warehouse

Hitachi Consulting
We make it happen. Better.



Azure SQL Data Warehouse

MPP - Logical Overview

Hitachi Consulting
We make it happen. Better.

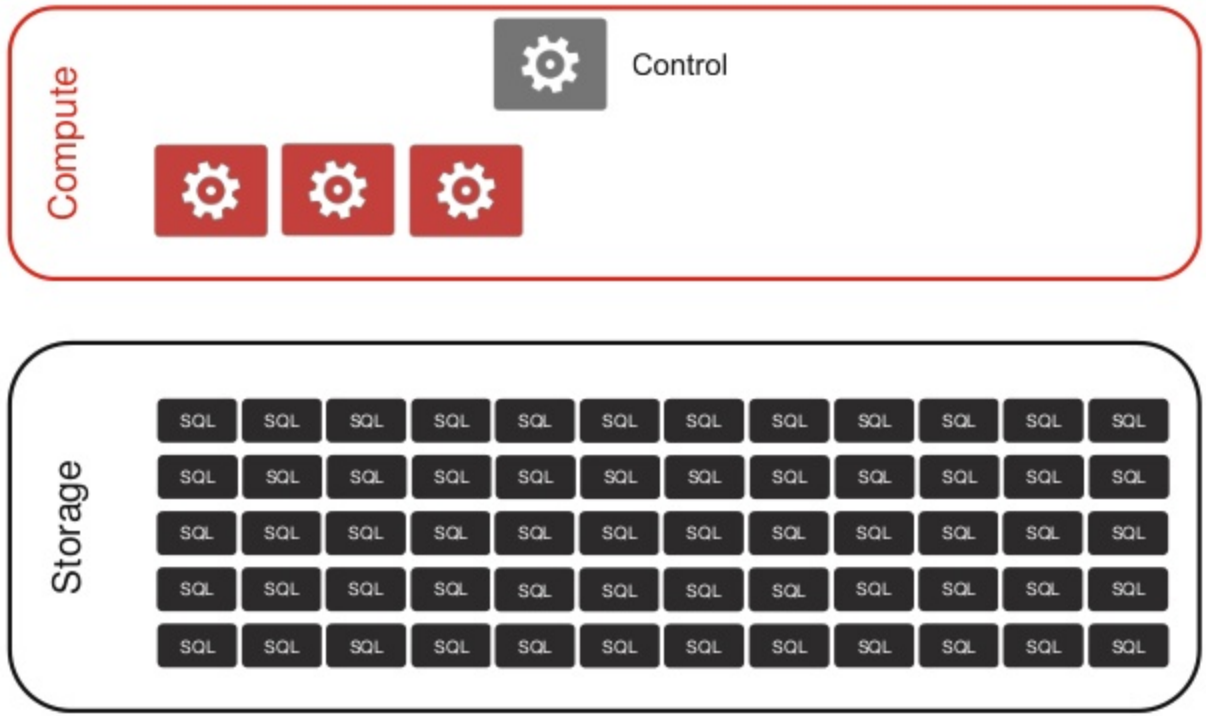
- 60 Data Distribution Nodes
- Each node contains a portion of data (distribution) for parallel processing
- Storage is de-coupled of compute

Azure SQL Data Warehouse

MPP - Logical Overview

Hitachi Consulting
We make it happen. Better.

- A scale-out distributed query engine



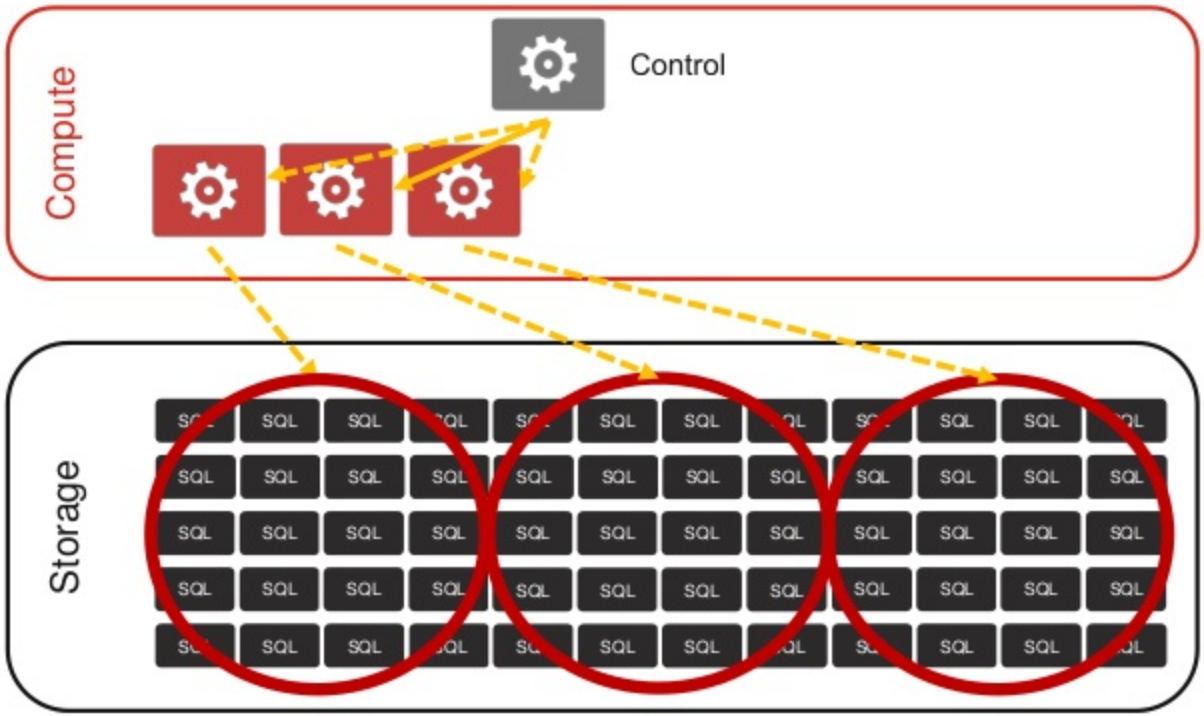
 |

Azure SQL Data Warehouse

MPP - Logical Overview

Hitachi Consulting
We make it happen. Better.

- A scale-out distributed query engine

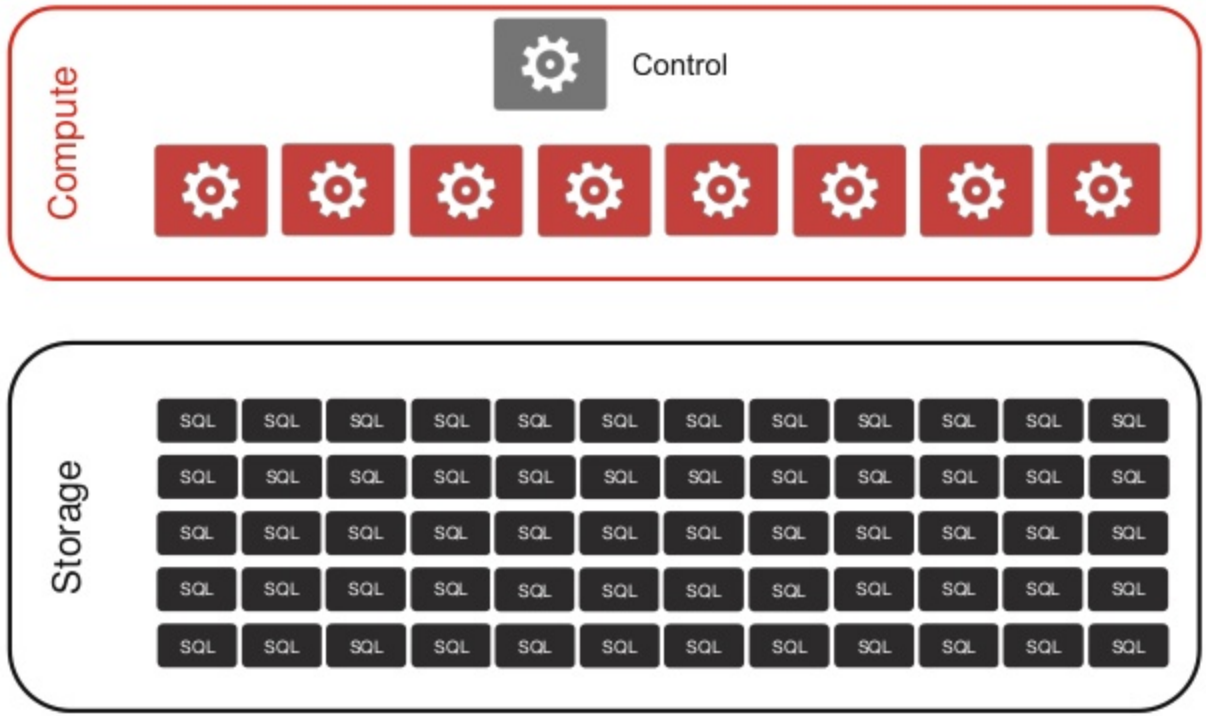


 |

Azure SQL Data Warehouse

MPP - Logical Overview

Hitachi Consulting
We make it happen. Better.



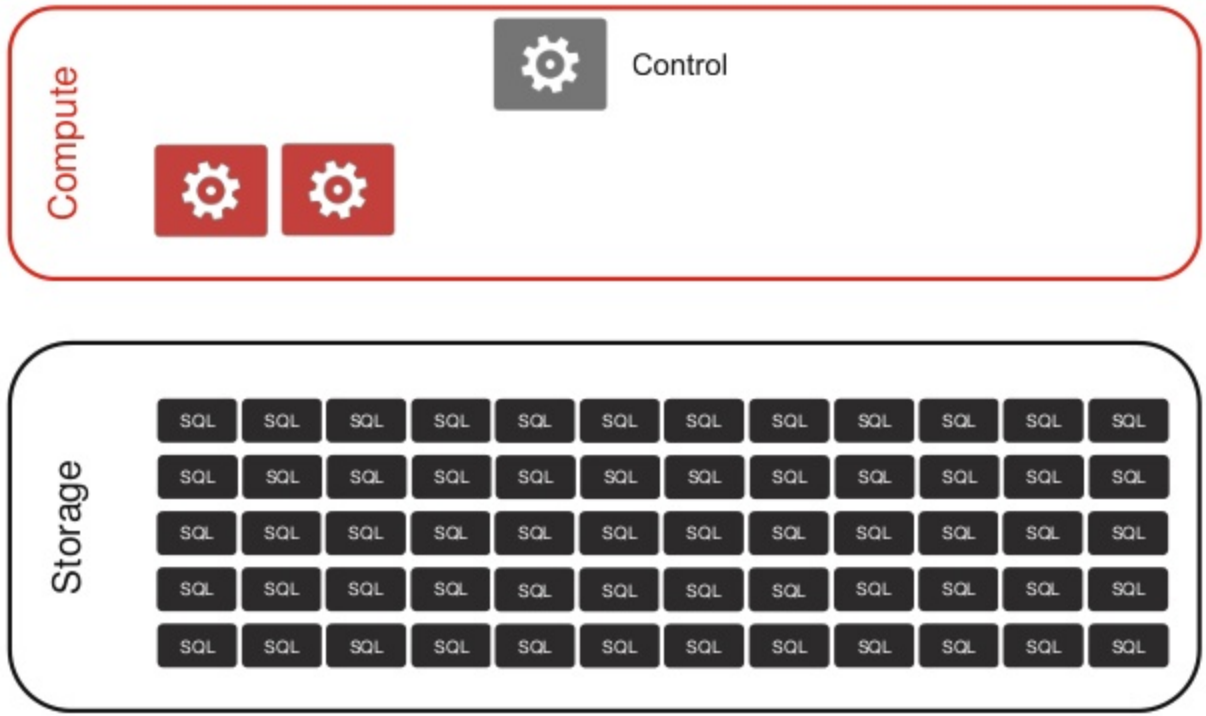
- A scale-out distributed query engine
- You get more compute capacity by increasing Data Warehouse Units (DWUs)

 |

Azure SQL Data Warehouse

MPP - Logical Overview

Hitachi Consulting
We make it happen. Better.



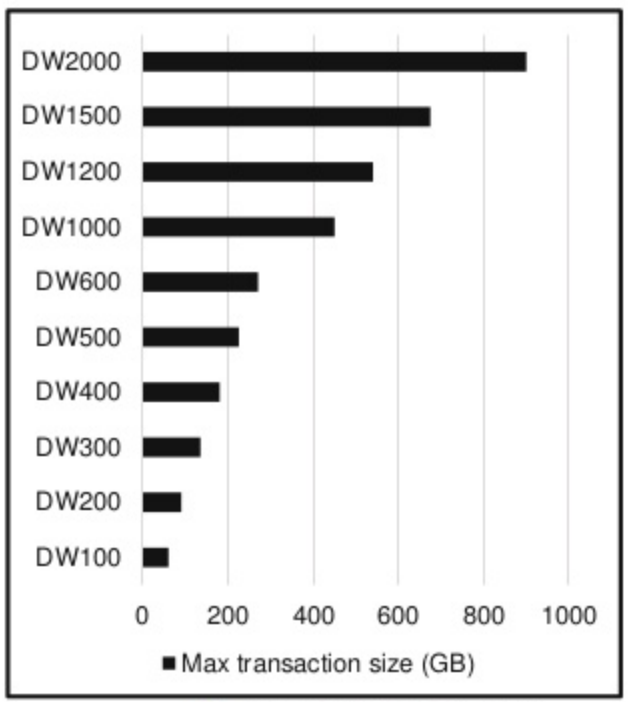
- A scale-out distributed query engine
- You get more compute capacity by increasing Data Warehouse Units (DWUs)
- Elastic – scale up/down as needed, and pay per usage (hourly-level)
- Scaling takes minutes as – compute is independent of data, thus no data re-distribution is needed.

Hitachi Consulting
We make it happen. Better.

Azure SQL Data Warehouse

Data Warehouse Units

RAM



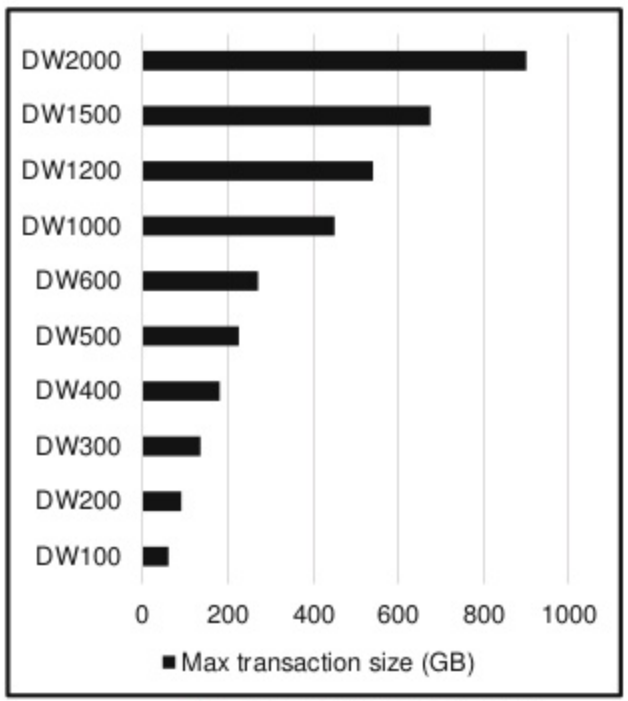
 |

Hitachi Consulting
We make it happen. Better.

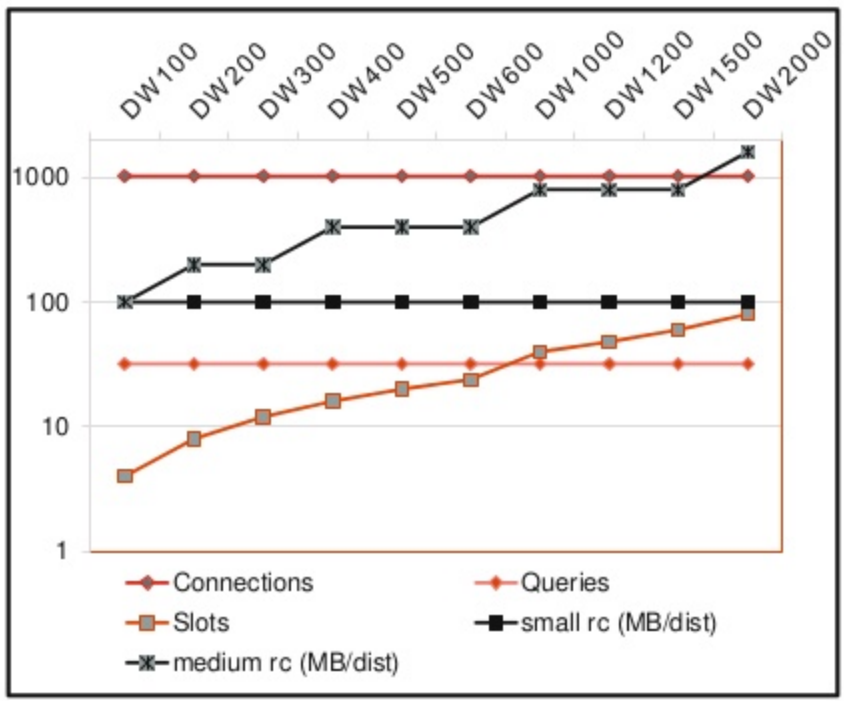
Azure SQL Data Warehouse

Data Warehouse Units

RAM



CPU



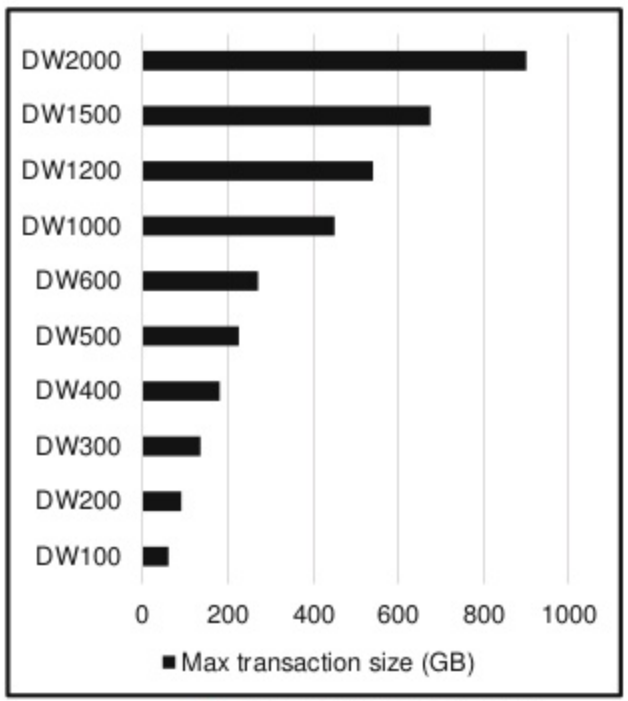
 |

Hitachi Consulting
We make it happen. Better.

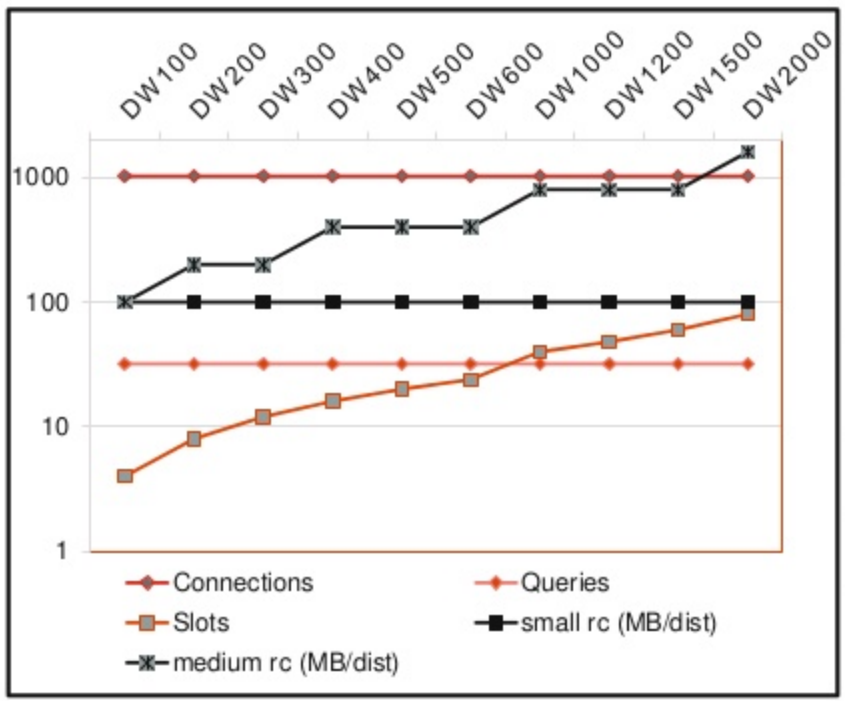
Azure SQL Data Warehouse

Data Warehouse Units

RAM



CPU



I/O

DWU	Readers	Writers
DW100	8	60
DW200	16	60
DW300	24	60
DW400	32	60
DW500	40	60
DW600	48	60
DW1000+	60	60

 |

Hitachi Consulting
We make it happen. Better.

Azure SQL Data Warehouse

Distribution vs. Partitioning

Distribution

- Aim to avoid **skewing** to maximize **balance parallel processing**.
  - if most of your data in one node, no parallel processing is gained.
- Aim to avoid **expensive data shuffling** across data nodes.
- Consider columns used in **JOIN** or **GROUP BY** for distribution keys to reduce shuffling
- Consider columns with **high cardinality** for distribution keys to avoid skewing
- **Round-Robin** distribution is the default - and a good compromise.
- Use **Replication** for small tables (yet to be available).

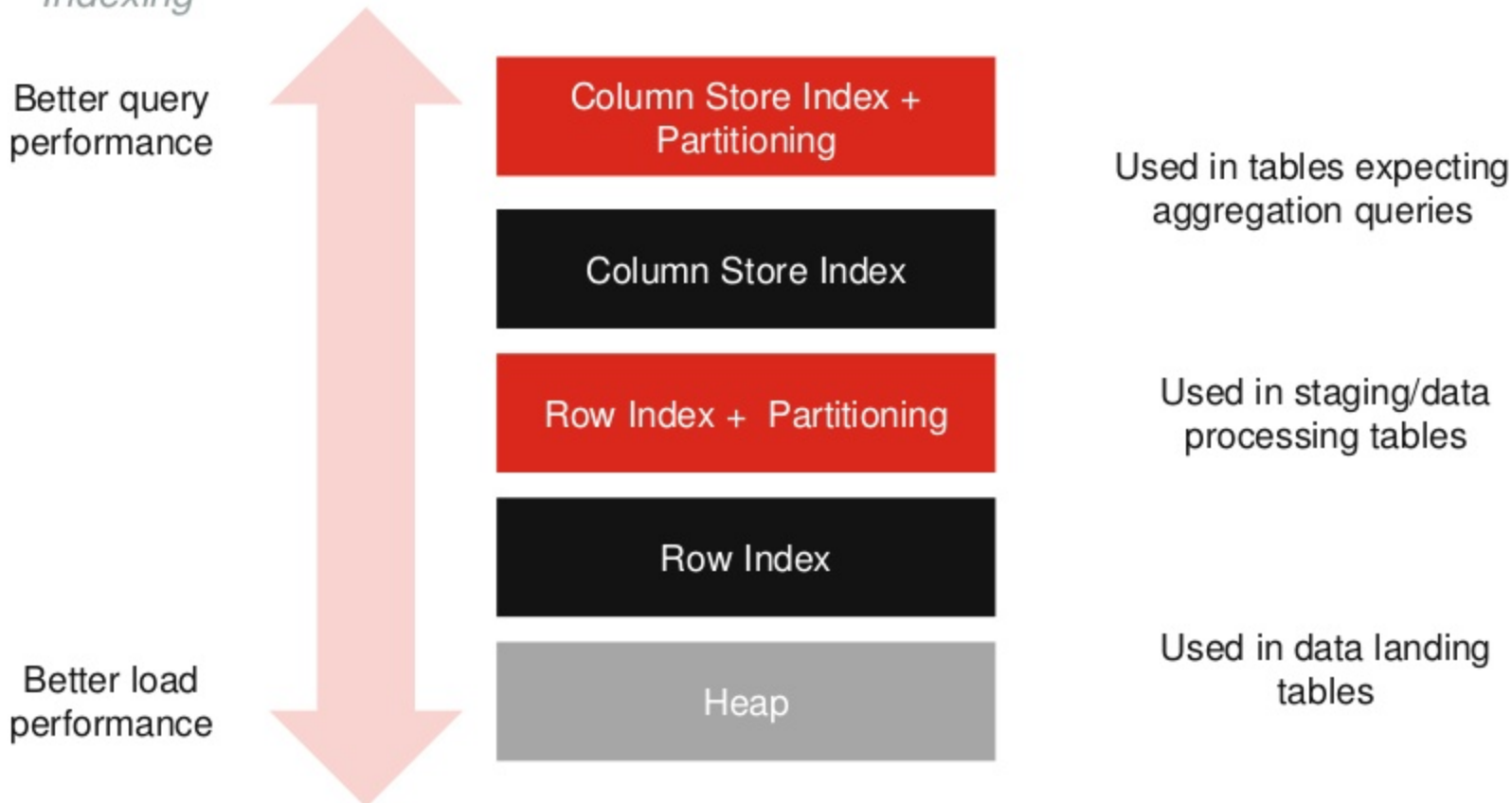
Partitioning

- In a given data node, aim to minimize the amount of data to be considered in the query
  - that is, **partition elimination**.
- Consider columns used in **WHERE** clauses for partition keys.

Hitachi Consulting
We make it happen. Better.

Azure SQL Data Warehouse

Indexing



 |

Hitachi Consulting
We make it happen. Better.

Azure SQL Data Warehouse

Data Model

- Capture elementary source data structure, main business entities and relationships
- Atomic, fine-grained, integrated, detail-oriented, historical, bi-temporal
- Provides a flexible, yet unified, data foundation allow agile business rules to construct new data products and information models.

Data Vault (Hubs, Links, Satellites)	3rd Normal form (3NF)	Dimensional Model (Facts, Dimension)
Dan Linstedt	Bill Inmon	Ralph Kimball

 |

Hitachi Consulting
We make it happen. Better.

Azure SQL Data Warehouse

PolyBase

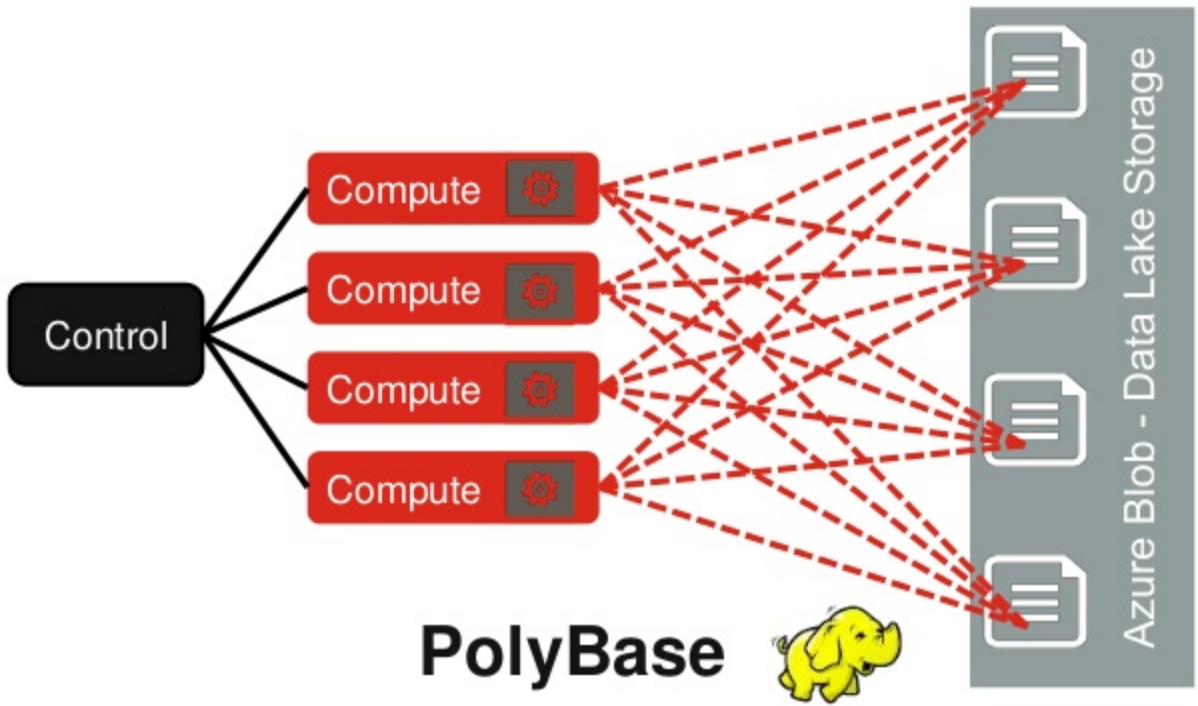
Hadoop-base technology (very similar to Hive)

Allow brining data in/out Azure SQL DW/SQL from/to Azure Blob Storage/Data Lake

Create External table (Schema-on-Read) – Identify file location and format

Parallel Data Movement – Direct connectivity between compute and data files (back door)

Best way to get data in and out of Azure SQL DW



 |

Azure SQL Data Warehouse

PolyBase

Hitachi Consulting
We make it happen. Better.

```
CREATE EXTERNAL DATA SOURCE MyAzureDataSource
WITH
(TYPE            = HADOOP
,LOCATION        =
'wasbs://[container]@accountname.blob.core.windows.net/path'
);


CREATE EXTERNAL FILE FORMAT MyTextFileFormat
WITH
(     FORMAT_TYPE        = DELIMITEDTEXT
     ,FORMAT_OPTIONS
     (
         FIELD_TERMINATOR = '|',
         STRING_DELIMITER = ',',
         DATE_FORMAT= 'yyyy-MM-dd',
         USE_TYPE_DEFAULT= TRUE
     )
     ,DATA_COMPRESSION  =
| 'org.apache.hadoop.io.compress.GzipCodec'
);
```

```
CREATE EXTERNAL TABLE [asb].[FactOnlineSales]
([ProductKey]    int      NOT NULL
,[StoreKey]      int      NOT NULL
,[DateKey]       int      NOT NULL
,[CustomerKey]   int      NOT NULL
,[PromotionKey]  int      NOT NULL
,[SalesQuantity] int      NOT NULL
,[UnitPrice]     money    NOT NULL
,[SalesAmount]   money    NOT NULL
)
WITH
(LOCATION='wasbs://filepath_or_directory'
,DATA_SOURCE     = MyAzureDataSource
,FILE_FORMAT     = MyTextFileFormat
,REJECT_TYPE     = VALUE
,REJECT_VALUE    = 0
,REJECT_SAMPLE_VALUE   = 1000
);
```


Hitachi Consulting
We make it happen. Better.

Azure SQL Data Warehouse

Data Processing

Avoid Singleton Inserts, Deletes, and Updates

Used Create Table As Select (CTAS) for processing data

Distribution

Partitioning

Indexing

Create and Update Statistics

Extract the data from the source, **L**oad the data, then perform the set-based **T**ransformations as SQL procedures to benefit from MPP - No SSIS data flow task!

Materialize external tables before processing

 |

Information Marts

 |

Hitachi Consulting
We make it happen. Better.

Information Marts

Role in the Enterprise Data Platform

Subject-Oriented

Owned by business departments

Contains data aggregations and calculations to answer pre-defined business questions

Azure SQL Database | Azure SQL 2016 VM | Analysis Service Model

Dimensional Model or pre-canned datasets

May contain a sandbox area for experimentation

Data is processed on the MPP platform (Azure SQL DW) and then loaded in the mart

Built on top of the single version of the truth in the EDW

Suitable for interactive query and high concurrency

 |

End-to-end Data Processing

Hitachi Consulting
We make it happen. Better.

Data Lake
<Azure Data Lake | Blob Storage>

Enterprise DW
<Azure SQL DW>

Information Mart
<Azure SQL DB | SQL Server 2016>

Big Data Processing (TEL)
<Data Lake Analytics | HDInsight>

Massively Parallel Processing (ELT)
<PolyBase + T-SQL>

Populate (ETL)
<Azure Data Factory | SSIS>

 |

Hitachi Consulting
We make it happen. Better.

Azure Data Factory

Hitachi Consulting
We make it happen. Better.

Azure Data Factory

Cloud data processing & movement services

A managed Azure cloud service for building & operating data processing and movements

Scalable

Reliable

Pay-as-you use

Compose, monitor & schedule data pipelines

Automatic cloud resource management

 |

Azure Data Factory

Concepts & Components

Hitachi Consulting
We make it happen. Better.

Data Factory

Linked Service

A connection information to a data store or a compute resource

Dataset

A pointer to the data object to be used as input or an output of an Activity

Pipeline

logical grouping of Activities with certain sequence & schedule

Azure Data Factory

Concepts & Components

Hitachi Consulting
We make it happen. Better.

Data Factory

Pipeline

Activity

actions to perform on data.
Input 0-N dataset(s) - Output 1-N dataset(s)

Linked Service

A connection information to a data store or a compute resource

Dataset

A pointer to the data object to be used as input or an output of an Activity

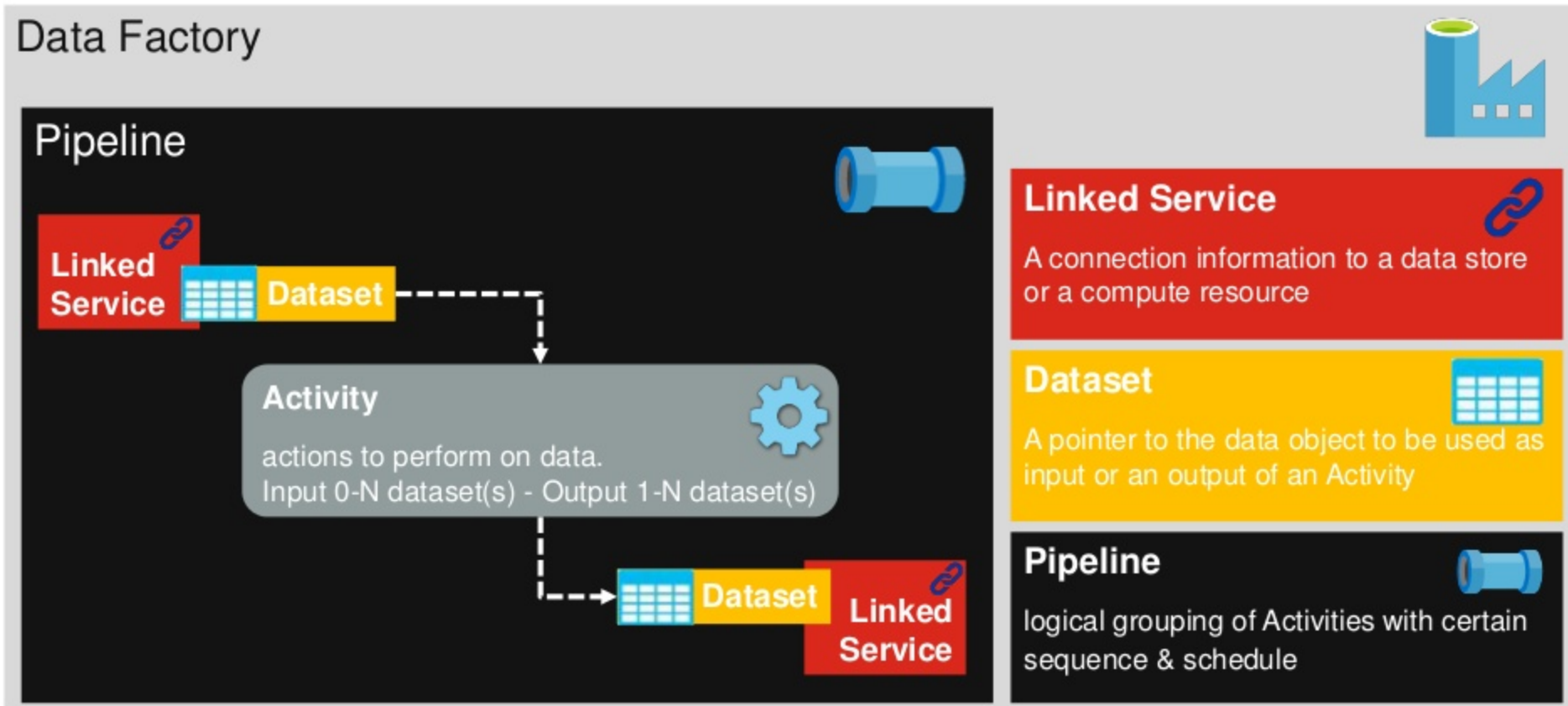
Pipeline

logical grouping of Activities with certain sequence & schedule

Azure Data Factory

Concepts & Components

Hitachi Consulting
We make it happen. Better.



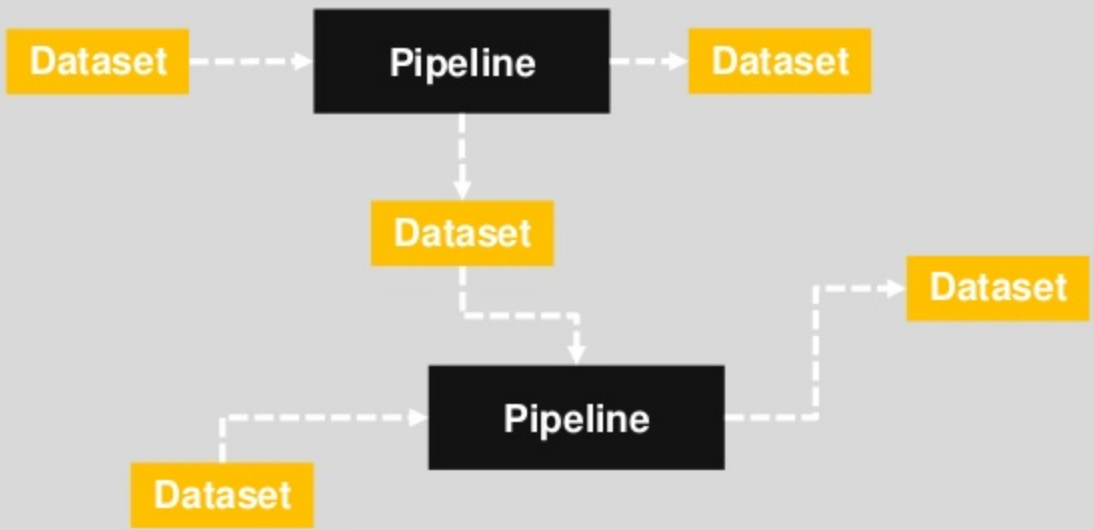
 |

Azure Data Factory

Concepts & Components

Hitachi Consulting
We make it happen. Better.

Data Factory



Linked Service

A connection information to a data store or a compute resource

Dataset

A pointer to the data object to be used as input or an output of an Activity

Pipeline

logical grouping of Activities with certain sequence & schedule

 |

Azure Data Factory

Concepts & Components

Hitachi Consulting
We make it happen. Better.

 |

Hitachi Consulting
We make it happen. Better.

Azure Data Factory

Features

Linked Service – Data

- Azure Blob
- Azure Table
- Azure SQL Database
- Azure SQL Data Warehouse
- Azure DocumentDB
- Azure Data Lake Store
- SQL Server
- ODBC data sources
- Hadoop Distributed File System (HDFS)

Linked Service - Compute

- Azure HDInsight On-Demand
  Azure Batch
- Azure Machine Learning
- Azure Data Lake Analytics
- Azure SQL DW (StoedProc)
- Azure SQL DB (StoedProc)
- SQL Server (StoredProc)

Activities

- Data Movement
  (copy Data from Source to sink)
- Data Transformation
  - Hive
  - Pig
  - MapReduce
  - Machine Learning
  - Stored Procedure
  - Data Lake U-SQL
  - Custom .NET

Data Gateway - Connect to an on-prem data source

Azure Data Factory

Example

Hitachi Consulting
We make it happen. Better.

Linked Service

```
{
  "name": "HDInsightOnDemandLinkedService",
  "properties": {
    "type": "HDInsightOnDemand",
    "typeProperties": {
      "version": "3.2",
      "clusterSize": 1,
      "timeToLive": "00:30:00",
      "linkedServiceName": "AzureStorageLinkedService1"
    }
  }
}
```

Input Dataset

```
{
  "name": "AzureBlobInput",
  "properties": {
    "type": "AzureBlob",
    "linkedServiceName": "AzureStorageLinkedService1",
    "typeProperties": {
      "fileName": "input.log",
      "folderPath": "adfgetstarted/inputdata",
      "format": {
        "type": "TextFormat",
        "columnDelimiter": ","
      }
    },
    "availability": {
      "frequency": "Month",
      "interval": 1
    },
    "external": true,        "policy": {}
  }
}
```

Output Dataset

```
{
  "name": "AzureBlobOutput",
  "properties": {
    "type": "AzureBlob",
    "linkedServiceName": "AzureStorageLinkedService1",
    "typeProperties": {
      "folderPath": "adfgetstarted/partitioneddata",
      "format": {
        "type": "TextFormat",
        "columnDelimiter": ","
      }
    },
    "availability": {
      "frequency": "Month",
      "interval": 1
    }
  }
}
```

Pipeline

```
{
  "name": "MyFirstPipeline",
  "properties": {
    "description": "My first Azure Data Factory pipeline",
    "activities": [
      {
        "type": "HDInsightHive",
        "typeProperties": {
          "scriptPath": "adfgetstarted/script/partitionweblogs.hql",
          "scriptLinkedService": "AzureStorageLinkedService1",
          "defines": {
            "inputtable": <storageaccounturl>",
            "partitionedtable": ,<folderurl>"
          }
        },
        "inputs": [
          {
            "name": "AzureBlobInput"
          }
        ],
        "outputs": [
          {
            "name": "AzureBlobOutput"
          }
        ],
        "policy": {
          "concurrency": 1,
          "retry": 3
        },
        "scheduler": {
          "frequency": "Month",
          "interval": 1
        },
        "name": "RunSampleHiveActivity",
        "linkedServiceName": "HDInsightOnDemandLinkedService"
      }
    ],
    "start": "2016-04-01T00:00:00Z",
    "end": "2016-04-02T00:00:00Z",
    "isPaused": false
  }
}
```

Azure NoSQL Data Stores

 |

Hitachi Consulting
We make it happen. Better.

What is NoSQL?

key attributes..

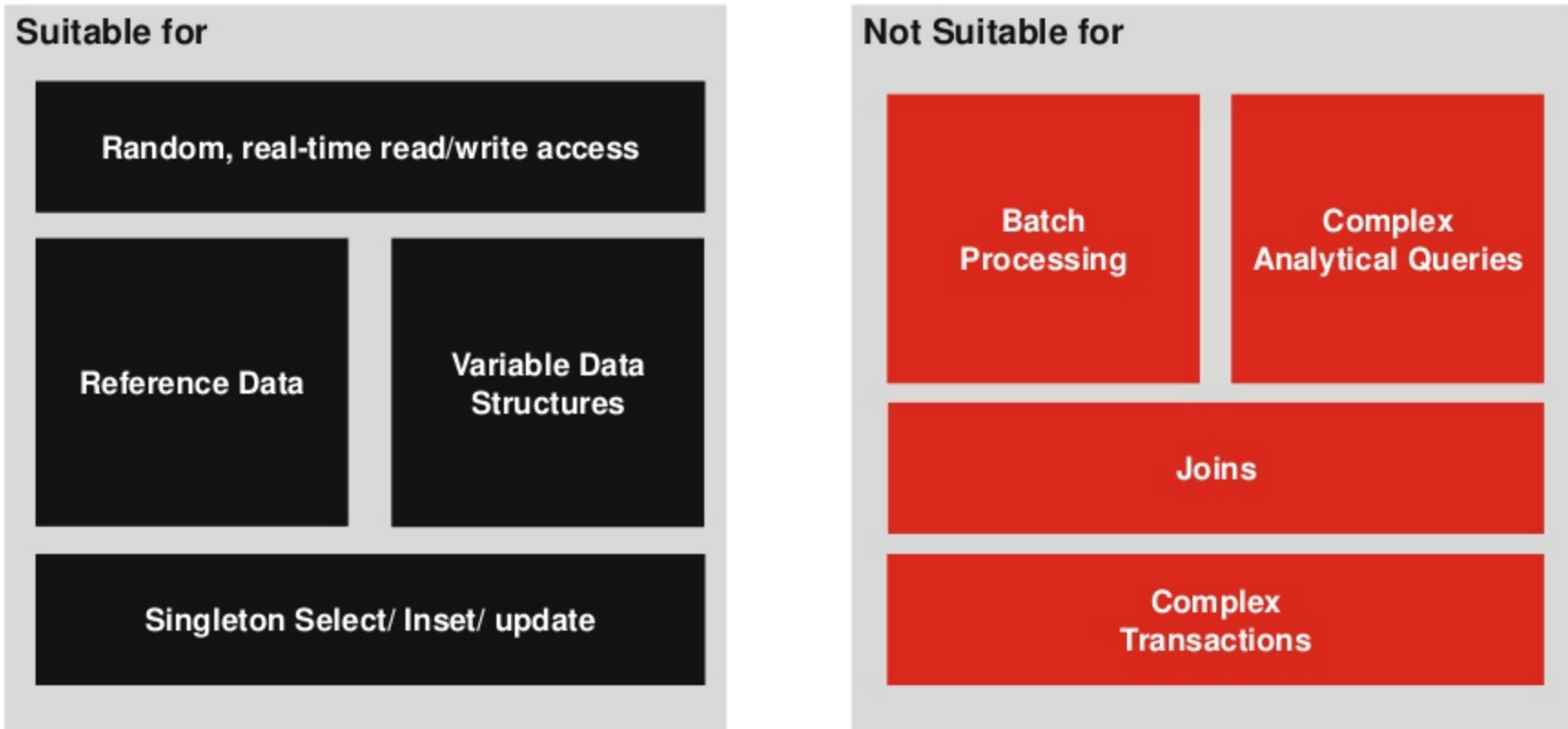
Scalable	Semi-structured	Complement RDBMS
Distributed	Non-relational	Flexible Schema
Fault-tolerant	Non-Transactional	Random, real-time read/write access

 |

Hitachi Consulting
We make it happen. Better.

Why NoSQL?

Suitability



 |

Hitachi Consulting
We make it happen. Better.

NoSQL Data Stores

Categories and breads

Key/Value Store

Column Family Store

Document Store

Graph Store

Memory Cache Store

Others

Hitachi Consulting
We make it happen. Better.

Azure NoSQL Technologies

Key/Value Store	▪ Flexible data structure ▪ Dictionary/ lookup ▪ Value can be anything
Column Family Store	▪ Column-oriented access ▪ Big Data with real-time read/write random access ▪ Extensible
Document Store	▪ Query-able data ▪ Objects (complex structure) in JSON, XML, etc. ▪ CRUD apps
Graph Store	▪ Social networks ▪ Fraud detection ▪ Relationship-heavy data
Memory Cache Store	▪ Non-durable data ▪ Fast access ▪ Hot data

Azure Table Storage

Azure Hbase on HDInsight

Azure DocumentDB

Microsoft Graph Engine (Trinity)

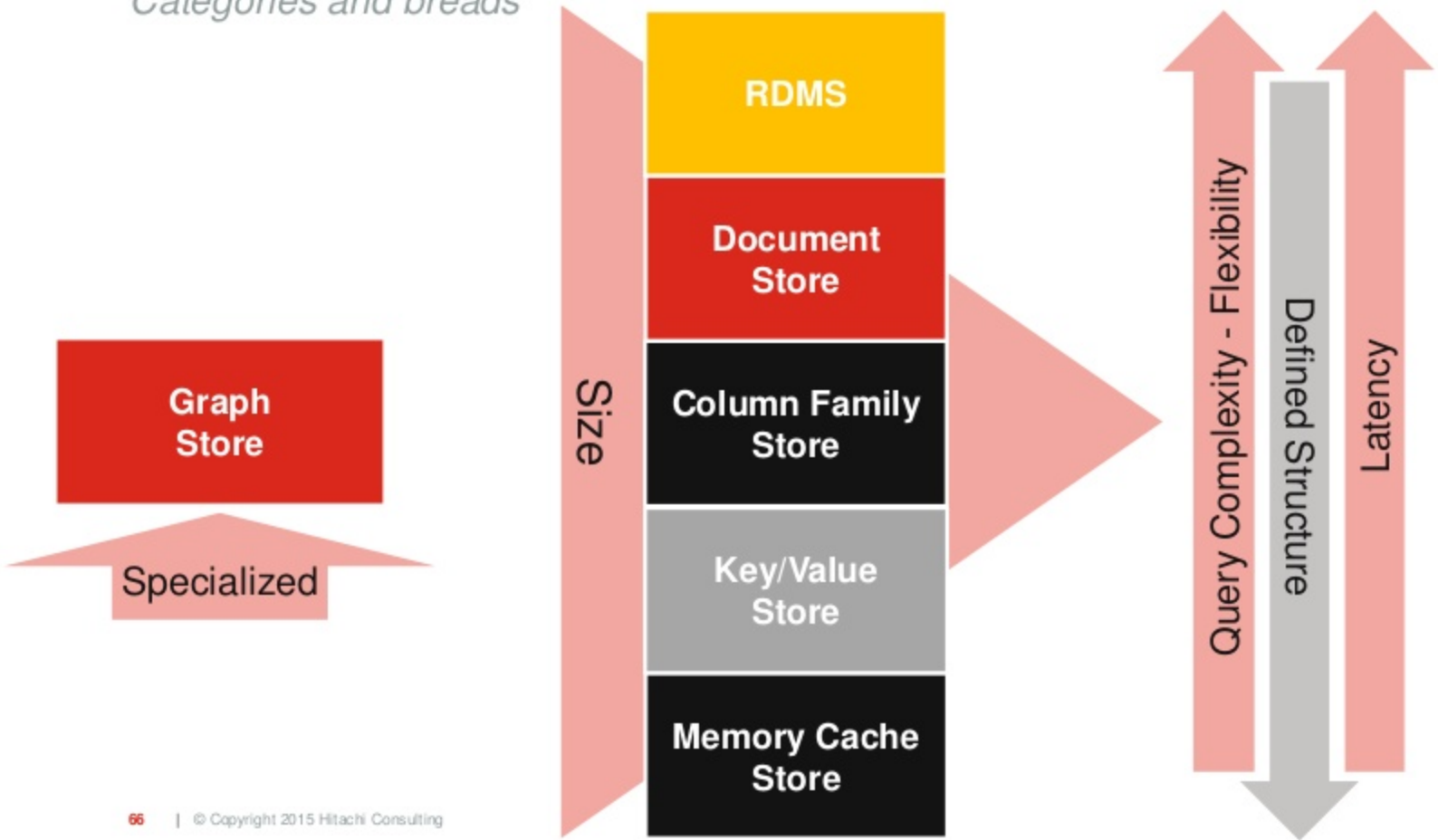
Azure Redis Cache

 |

Hitachi Consulting
We make it happen. Better.

NoSQL Data Stores

Categories and breads

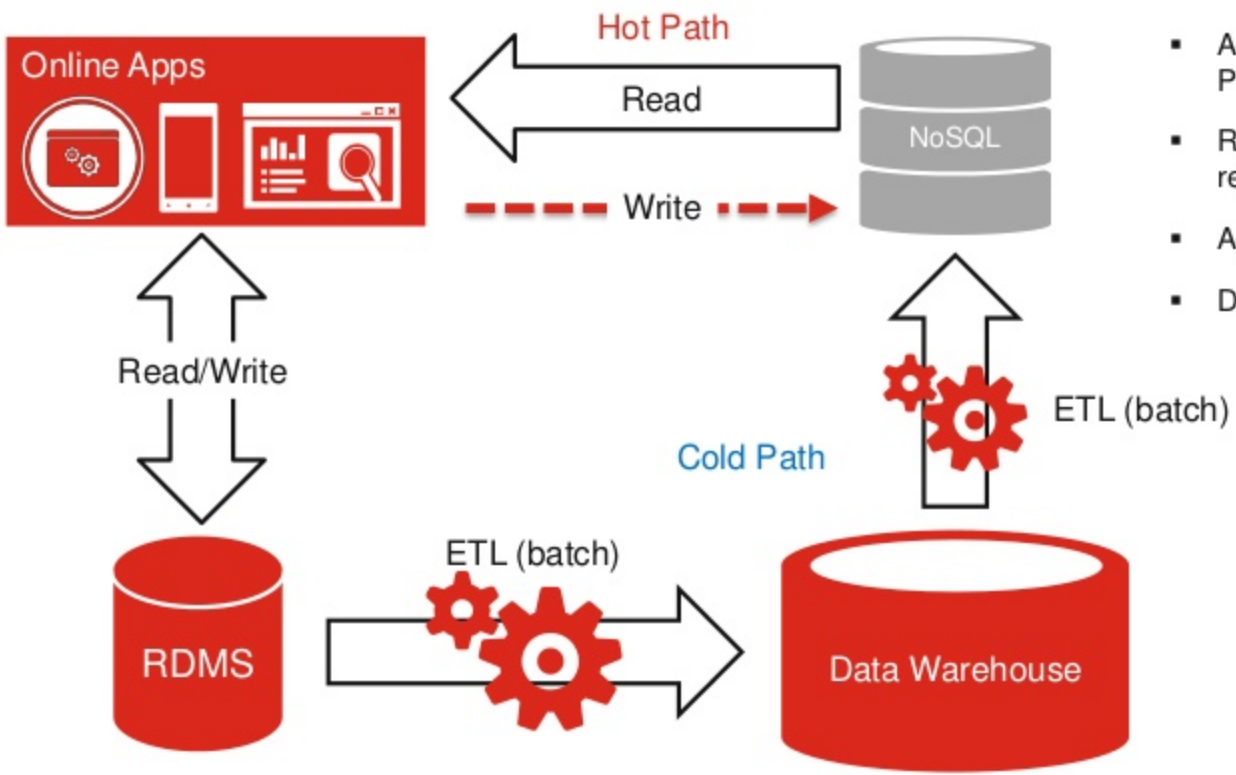


 |

NoSQL Usage Patterns

A common scenario…

Hitachi Consulting
We make it happen. Better.



- Data is Extracted, Transformed, loaded from OLTP to EDW
- Aggregations, KPIs, and scores are computed via Batch Processing
- Results are populated to a NoSQL data store for reference use in apps
- App hot read - ETL cold write
- Document & Graph stores

E.g., Single Customer View:

- Customer Matching, customer KPIs, segment assignment, and propensity scoring are performed as batch processing in DW
- The output goes to NoSQL to be used for real-time recommendation, campaigning, targeted advertising, etc.

 |

Hitachi Consulting
We make it happen. Better.

Real-time stream processing on Microsoft Azure

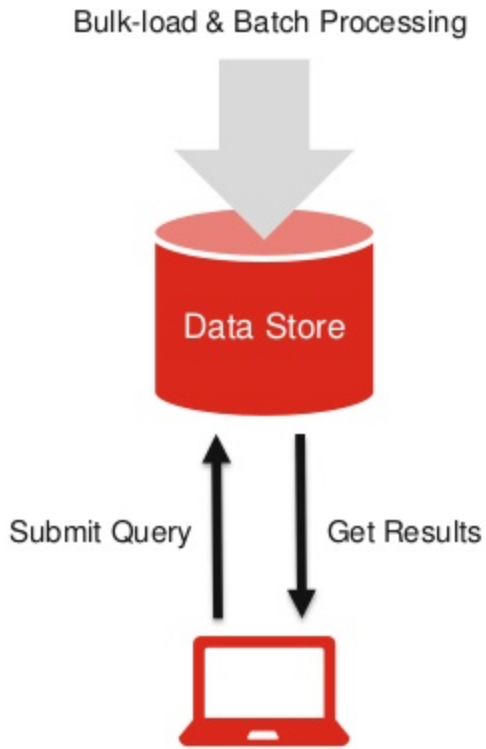
 |

Hitachi Consulting
We make it happen. Better.

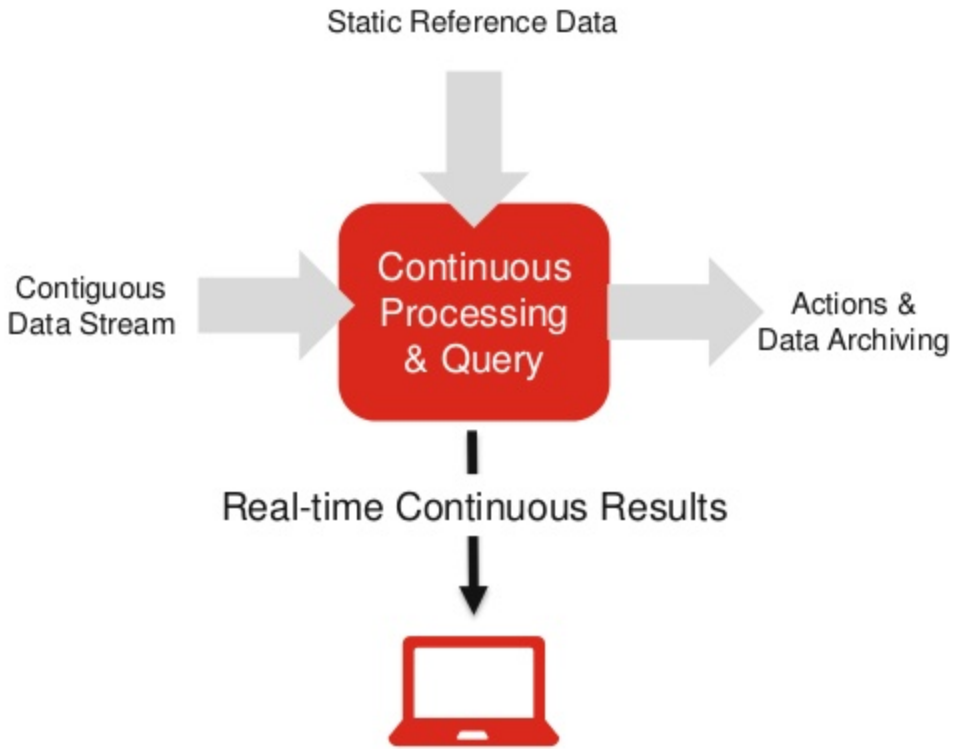
What is Event & Stream Processing?

Data at rest vs. Data in motion

Traditional – Working with data at rest



Real-time – Working with data at motion

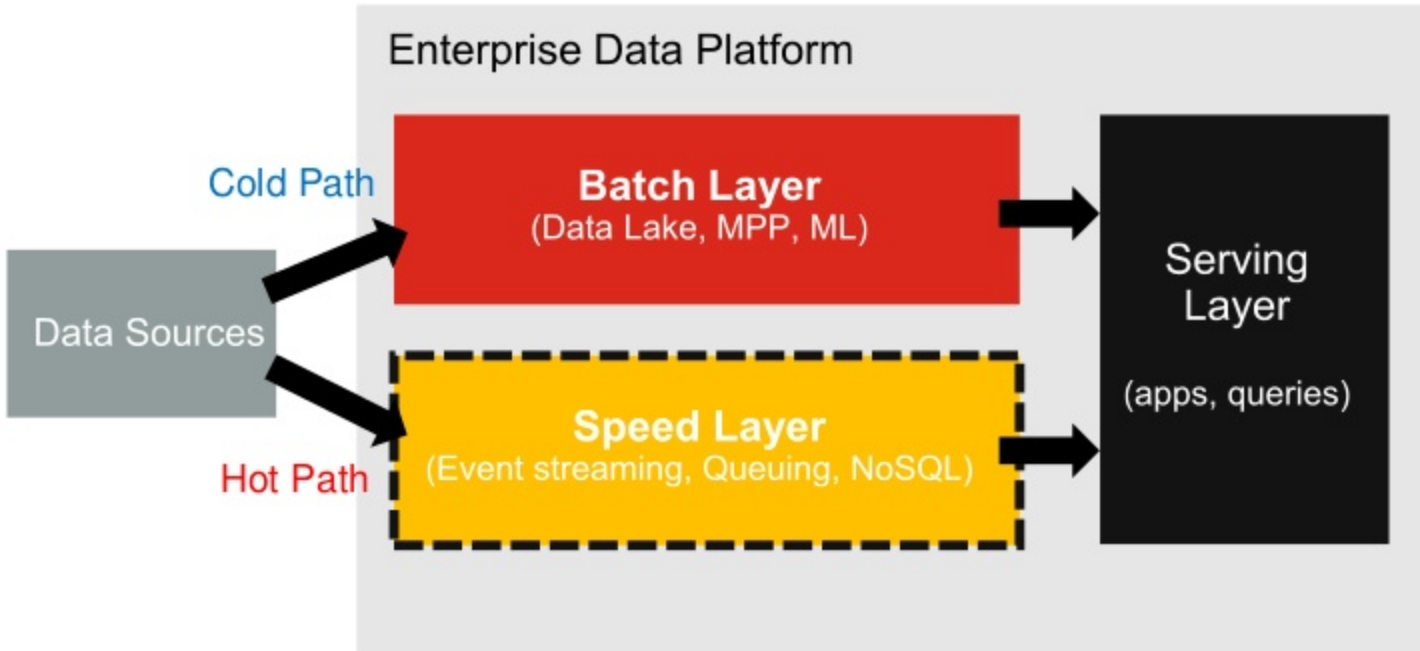


 |

Hitachi Consulting
We make it happen. Better.

Lambda Architecture

The speed layer and stream processing



 |

Hitachi Consulting
We make it happen. Better.

Events & Stream Processing Architecture

Azure Tools & Technologies

Event Triggers

Applications

Devices

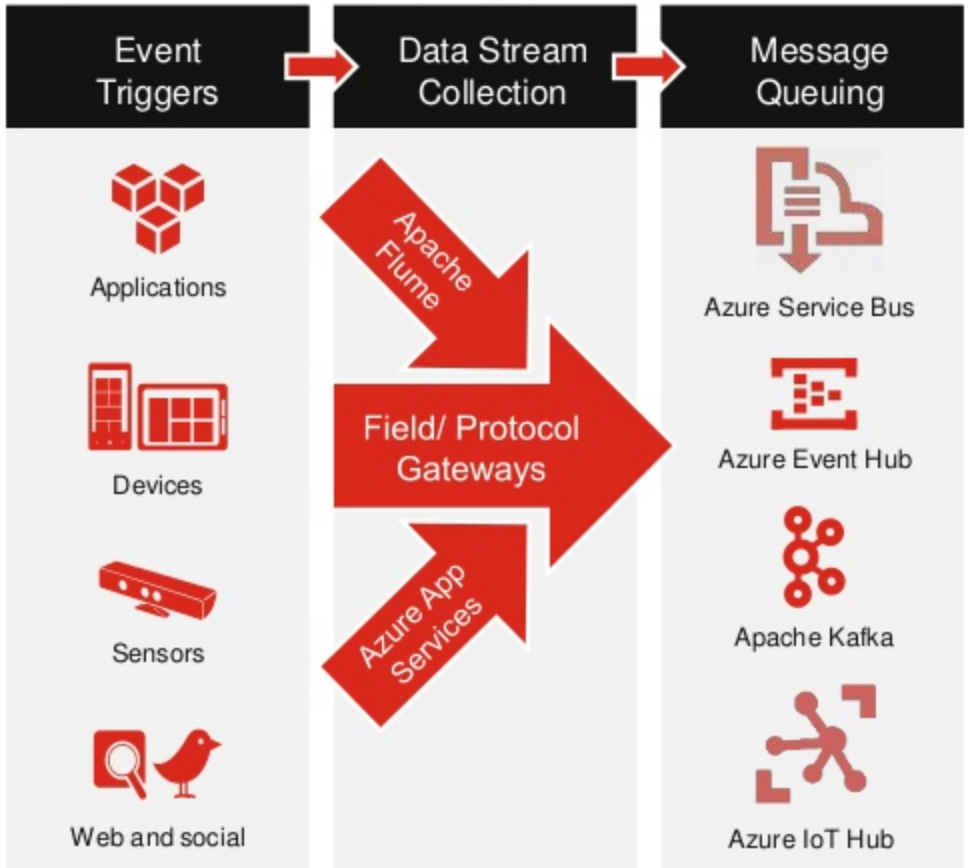
Sensors

Web and social

Hitachi Consulting
We make it happen. Better.

Events & Stream Processing Architecture

Azure Tools & Technologies

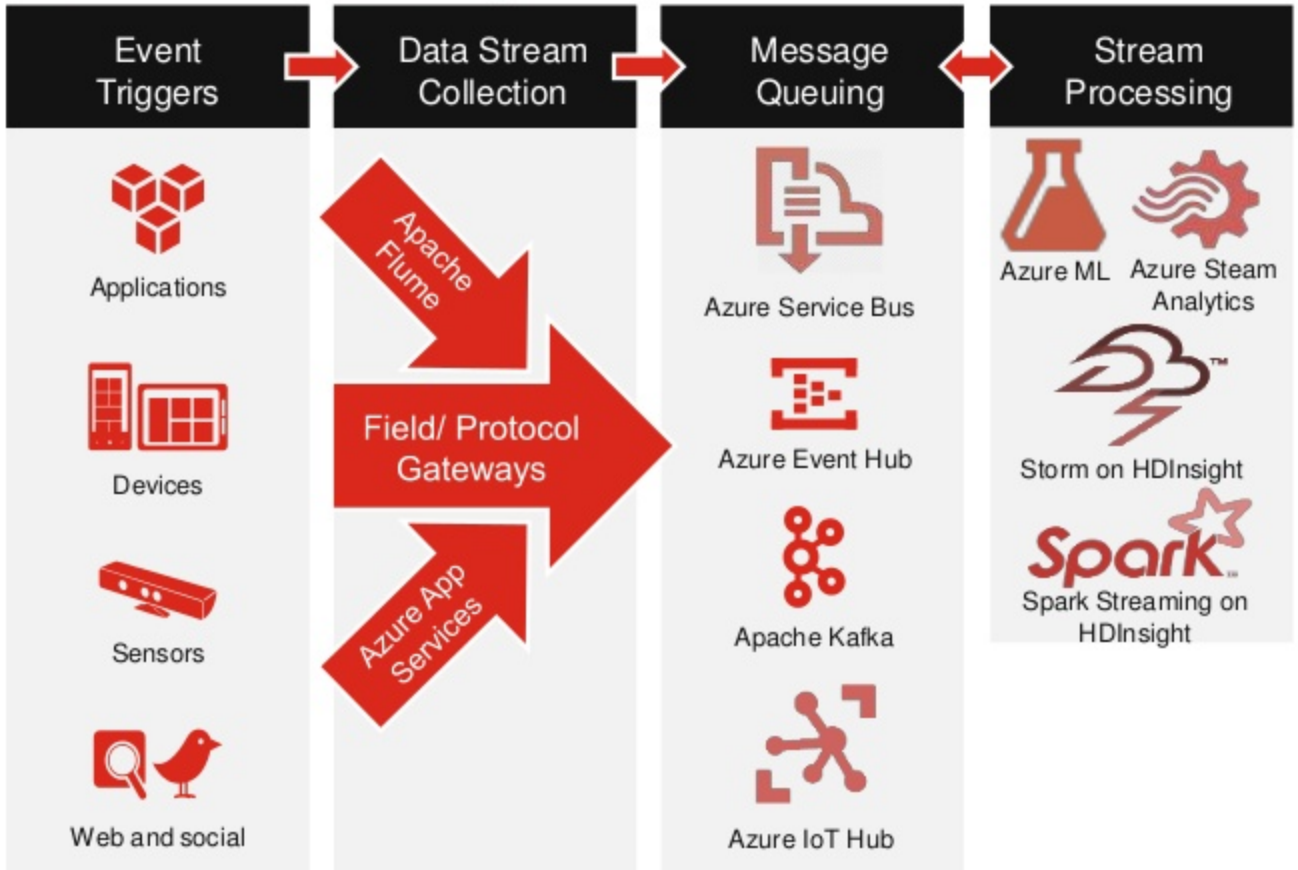


 |

Hitachi Consulting
We make it happen. Better.

Events & Stream Processing Architecture

Azure Tools & Technologies

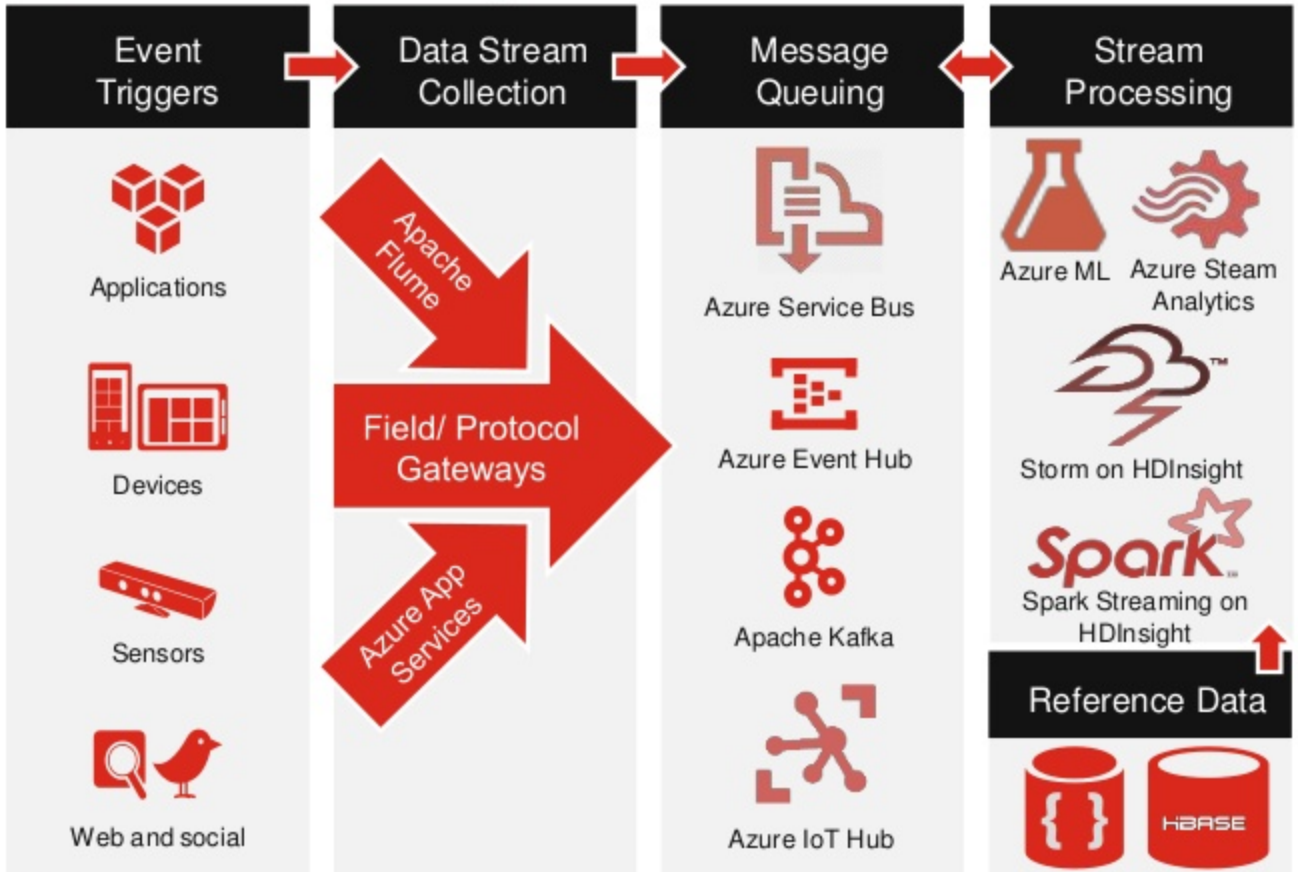


 |

Hitachi Consulting
We make it happen. Better.

Events & Stream Processing Architecture

Azure Tools & Technologies

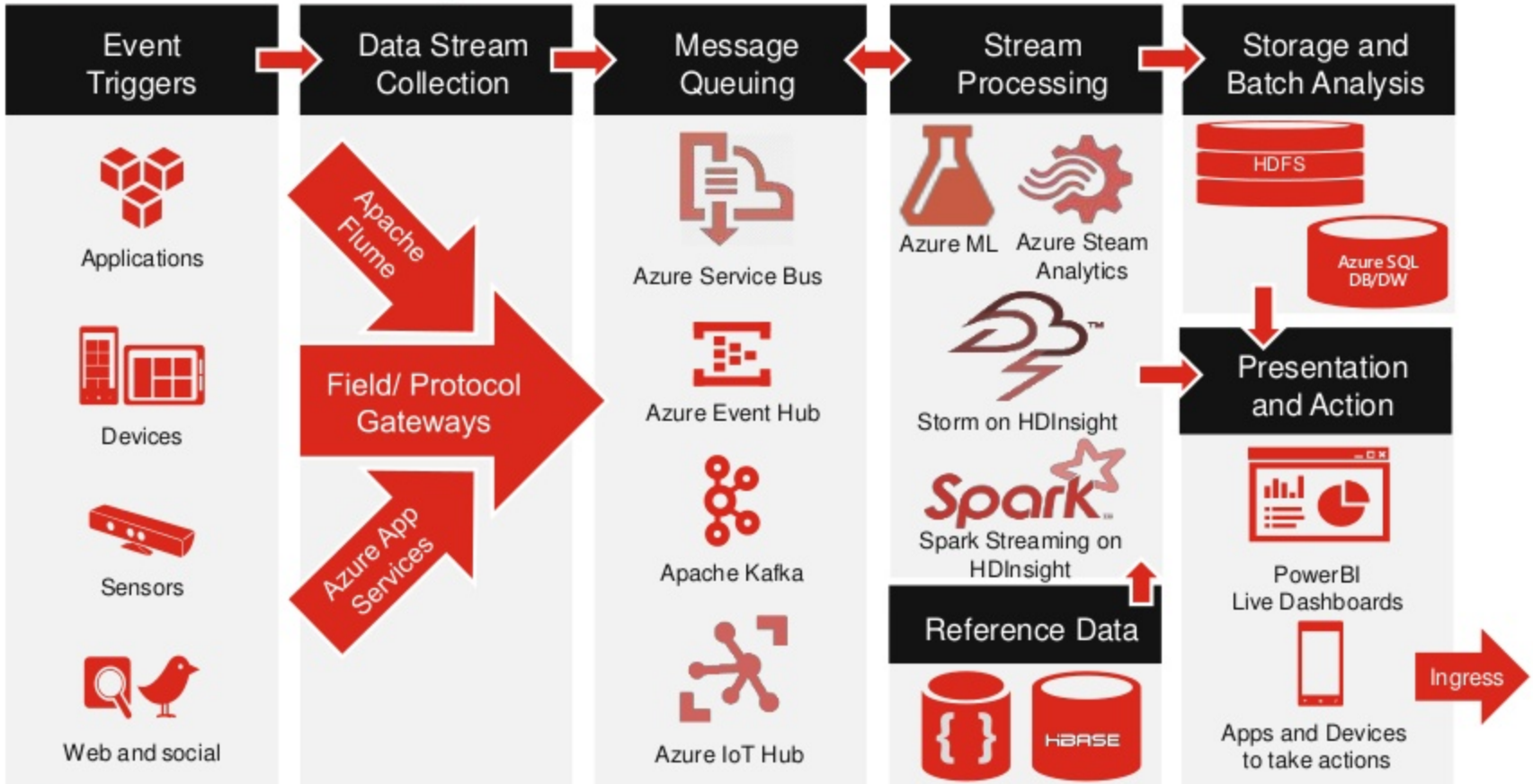


 |

Hitachi Consulting
We make it happen. Better.

Events & Stream Processing Architecture

Azure Tools & Technologies



 |

Events & Stream Processing Architecture

Hitachi Consulting
We make it happen. Better.

- Devices, Websites, and Apps that continuously produce data streams
- Listen to, collection, and transfer in-bound events
- De-couples data consumers from data producers
- Reliable, distributed fault-tolerant, high-throughputs short-tem storage
- Aggregate / filter / join incoming event streams
- Temporal engine for analysing data across time-series windows
- High throughputs, random access data store to support processing
- Usually NoSQL data stores
- Store processed/ aggregated/ filtered data (SQL/NoSQL)
- Consolidate and store raw data into files for batch analysis (DFS)
- Rich interactive visualizations for real-time data analysis
- Application integration for process automation

Azure Stream Analytics

Hitachi Consulting
We make it happen. Better.

A PaaS real-time **complex event processing** (CEP) on Microsoft Azure

Fully-managed real-time processing

- Intake millions of events per second
- Processing on continuous streams of data
- Reference data lookup
- Output to live dashboards and data sores

Mission Critical Reliability

- Guaranteed events delivery
- Preserves event order pre-device basis
- Guaranteed business continuity
- Auto-recovery from failures

No challenges with Scale

- Elasticity for scale up or scale down
- Distributed, scale-out architecture
- Pay only for the resources you use

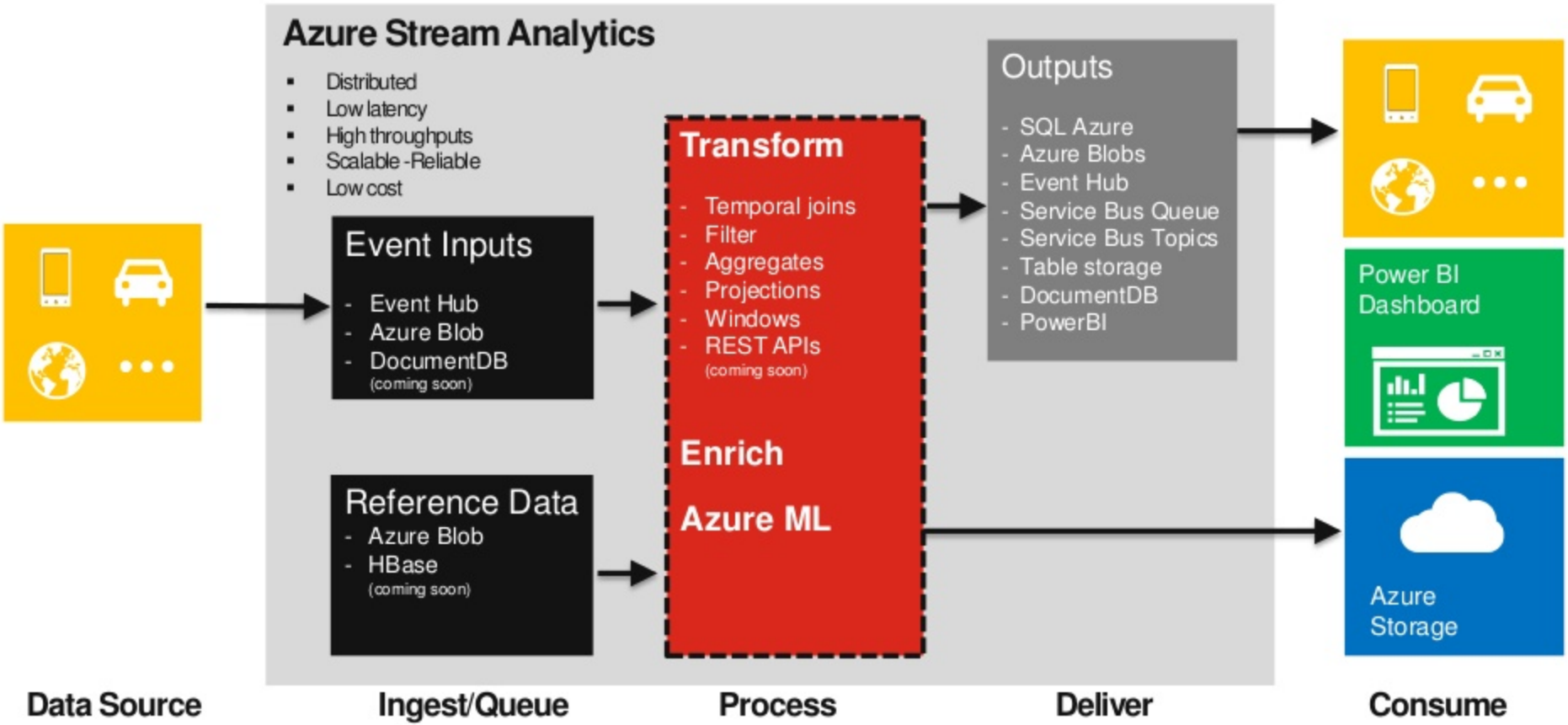
Rapid Development & Deployment

- SQL-like Language
- Built-in temporal semantics
- Up and running in a few clicks
- Scheduling and Monitoring

 |

Azure Stream Analytics

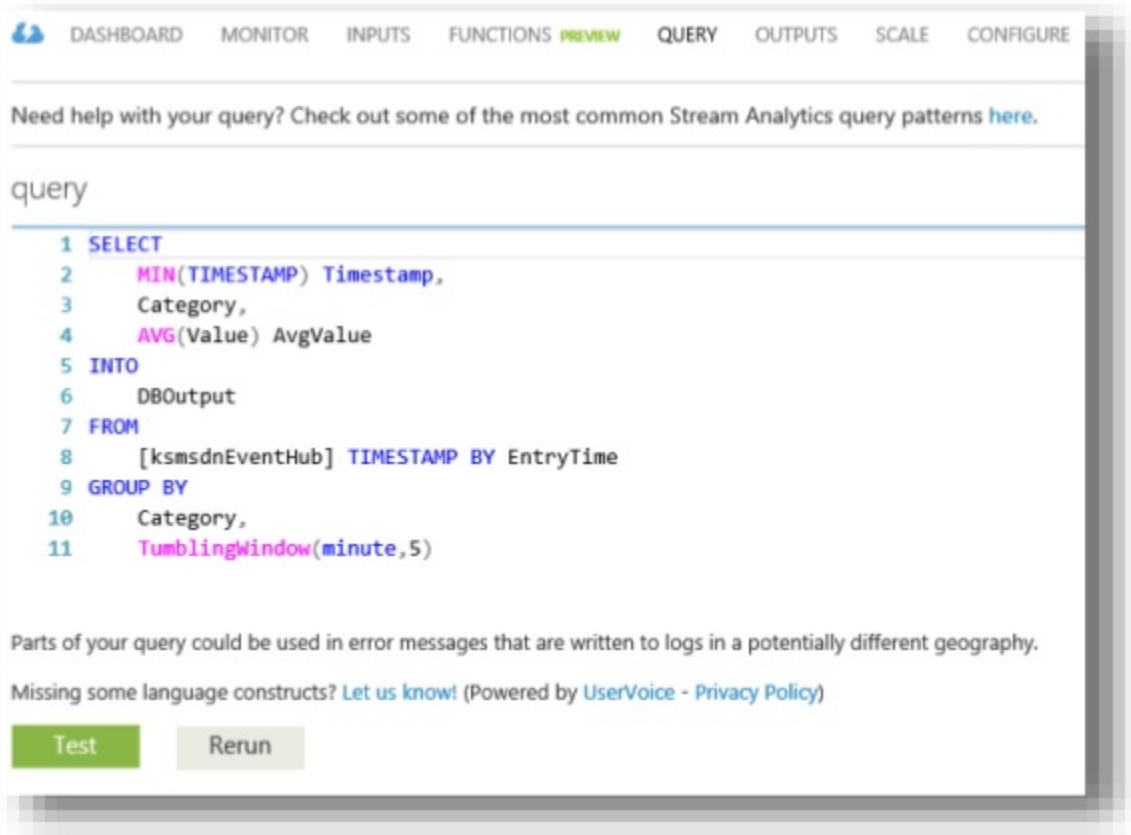
Hitachi Consulting
We make it happen. Better.



 |

Azure Stream Analytics

Stream Analytics Query



 |

Hitachi Consulting
We make it happen. Better.

Stream Analytics Query Language

Windowing Functions

- In **data streams**, a common requirement is to perform aggregation (max, min, sum, count, etc.) over messages that **arrive within a specified period of time** (window) - **to detect events.**
- Each **Group By** requires a **windowing function**
- Each window operation **outputs a single event** at the end of the window
- All windows have a **fixed length**

Tumbling window

Aggregate per time interval

Hopping window

Schedule overlapping windows

Sliding window

Windows constant re-evaluated

 |

Hitachi Consulting
We make it happen. Better.

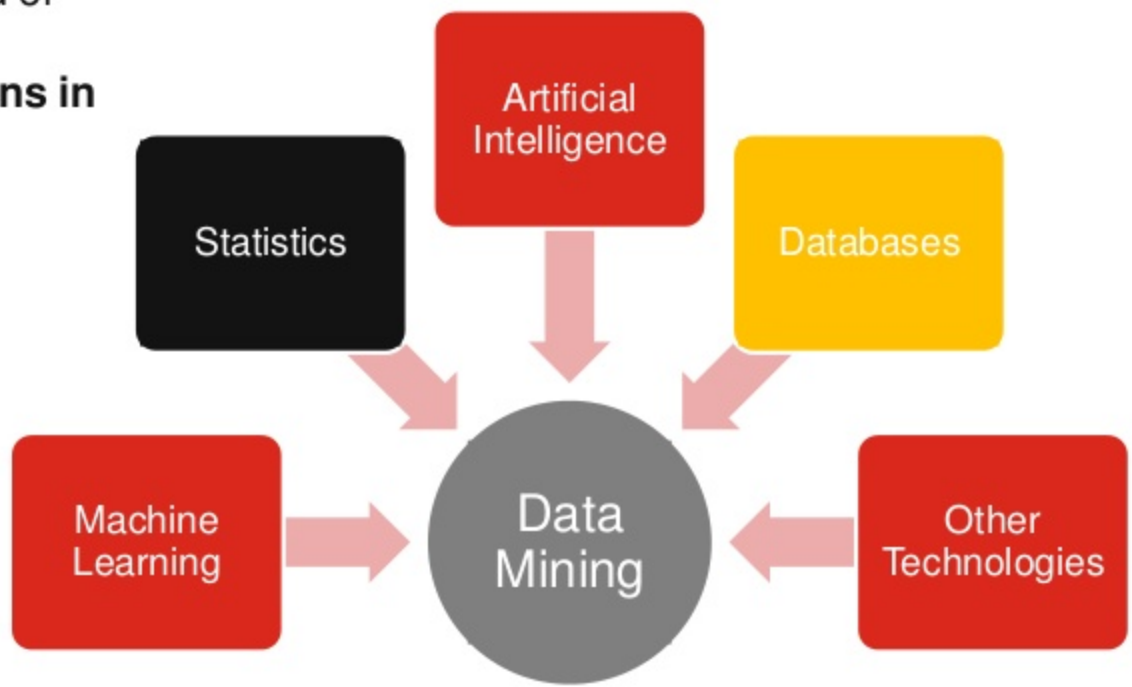
Data Science and Machine Learning on Microsoft Azure

Hitachi Consulting
We make it happen. Better.

Data Science and Machine Learning

What?

"Data mining, an interdisciplinary subfield of computer science, is the computational process of **automatic discovering patterns in large data sets"**



Other Related Technologies:

- Visualization
- Big Data
- High Performance Computing
- Cloud Computing
- Others..

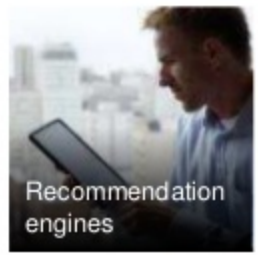
 |

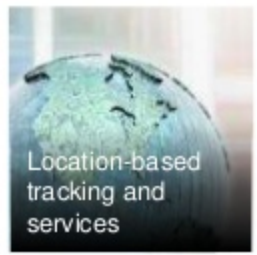
Data Science and Machine Learning

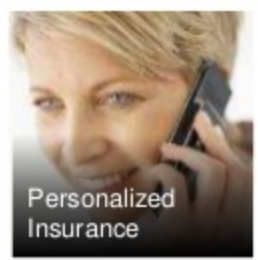
Why?

Hitachi Consulting
We make it happen. Better.

Machine learning & predictive analytics are core capabilities that are needed throughout your business



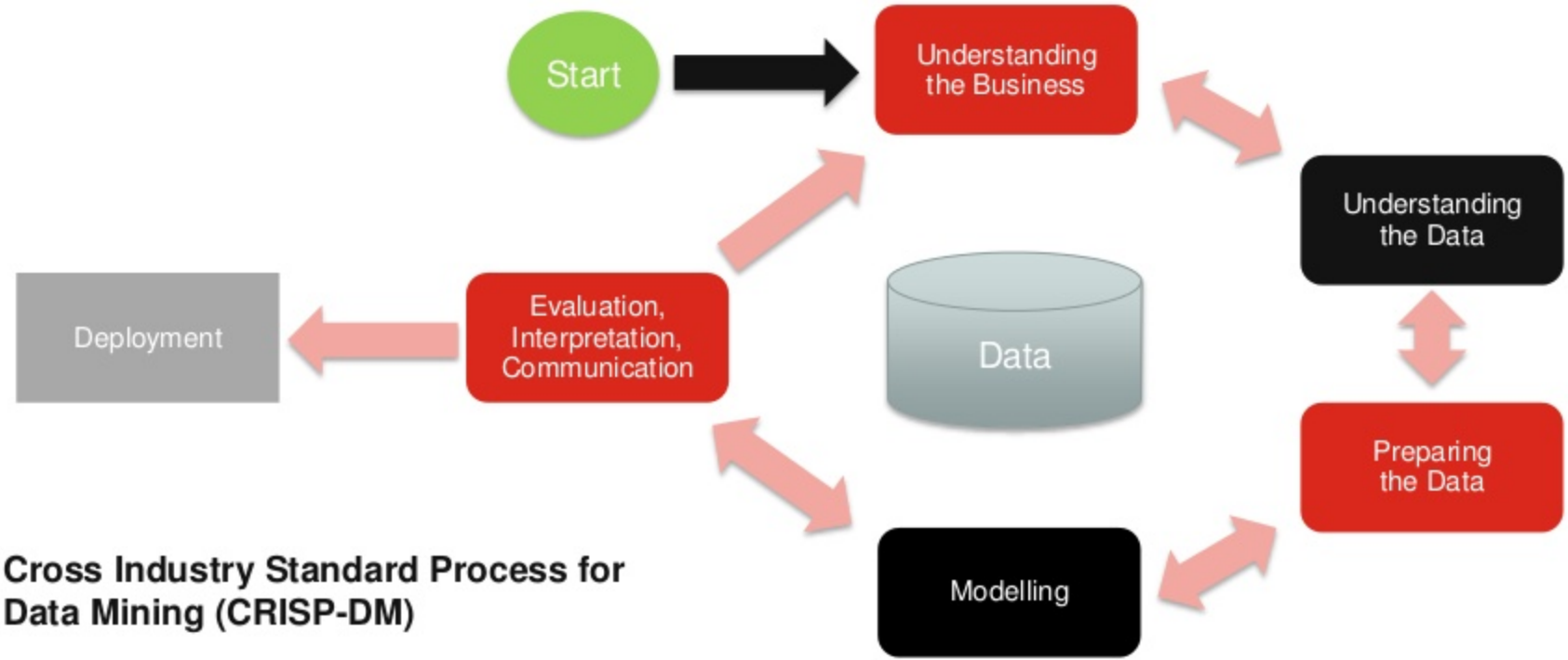




Hitachi Consulting
We make it happen. Better.

Data Science and Machine Learning

How?



Cross Industry Standard Process for Data Mining (CRISP-DM)

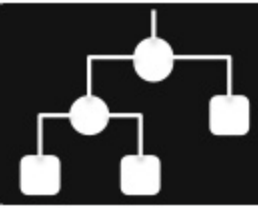
Hitachi Consulting
We make it happen. Better.

Data Science and Machine Learning

How?

Classification Learning

Build a model that can predict the target class of an input case



Time Series Analysis

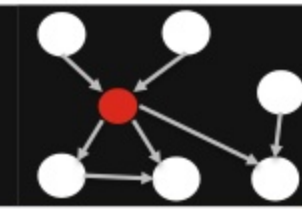
Analysis of temporal data to forecast future values

Regression Modeling

Build a model that can estimate the response value given an input case

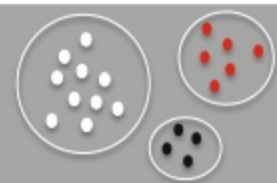
Probabilistic Modeling & Estimation

Compute the probability of an event to occur



Cluster Analysis

Discover natural groupings within the data points



Similarity Analysis

Identify similar cases to a given input case

Association Rule Discovery

Extract frequent patterns present in the data

```
IF .. AND .. AND ..
THEN A
ELSE IF .. AND ..
THEN C
ELSE IF .. AND ..
THEN B
..
```

Collaborative Filtering

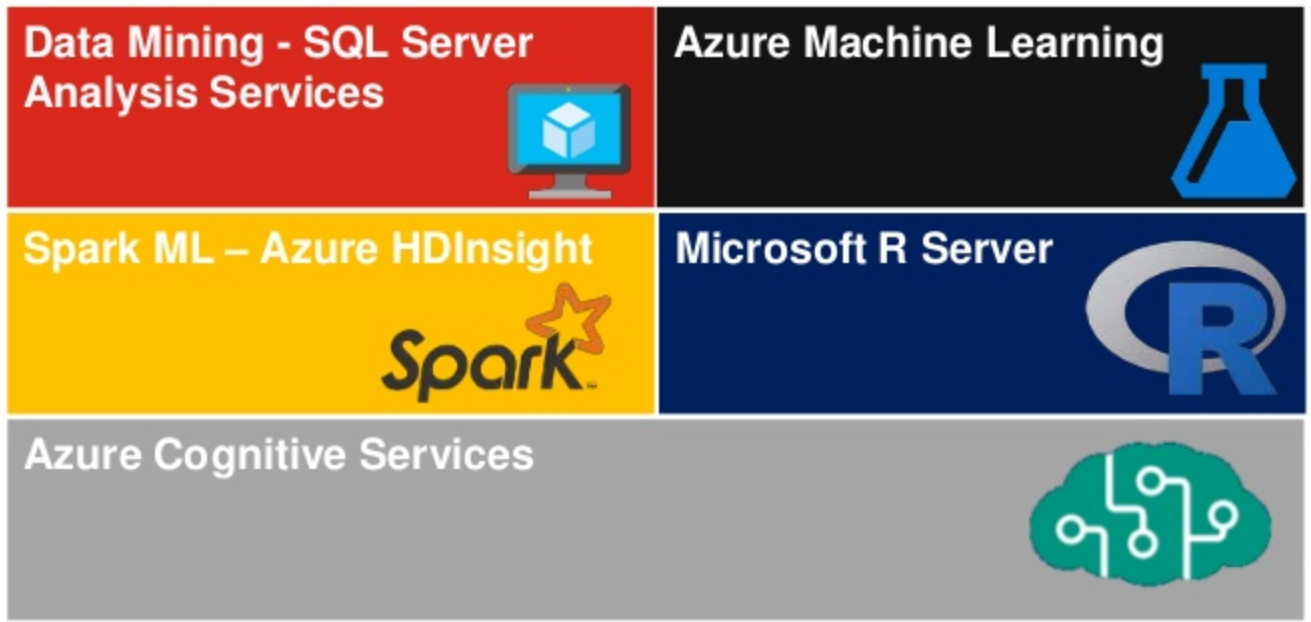
Filtering of information using techniques involving collaboration viewpoints

 |

Hitachi Consulting
We make it happen. Better.

Data Science and Machine Learning

Microsoft tools and technologies



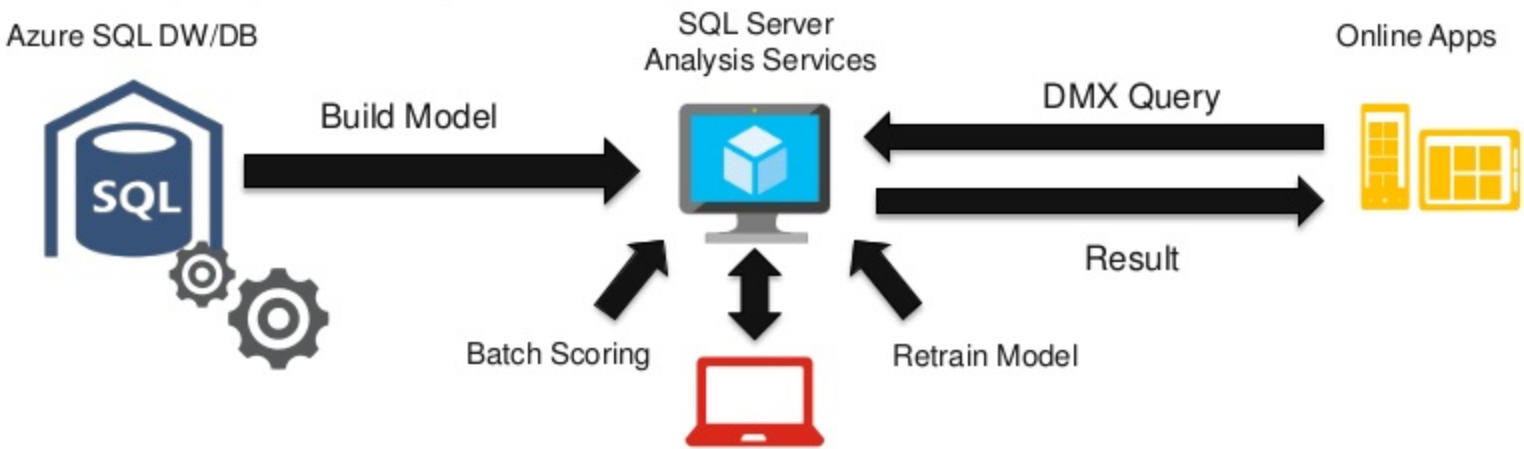
 |

Hitachi Consulting
We make it happen. Better.

Data Science and Machine Learning

Microsoft Analysis Services

- SQL Server Analysis Services (SQL Server on a VM, no PaaS)
- Process data from many OLEDB and ODBC data sources
- Easy to **build, interpret, deploy, and productionize**
- Limited extensibility
- Custom APIs can be built to use constructed models (DMX)
- Limited pre-processing functionality

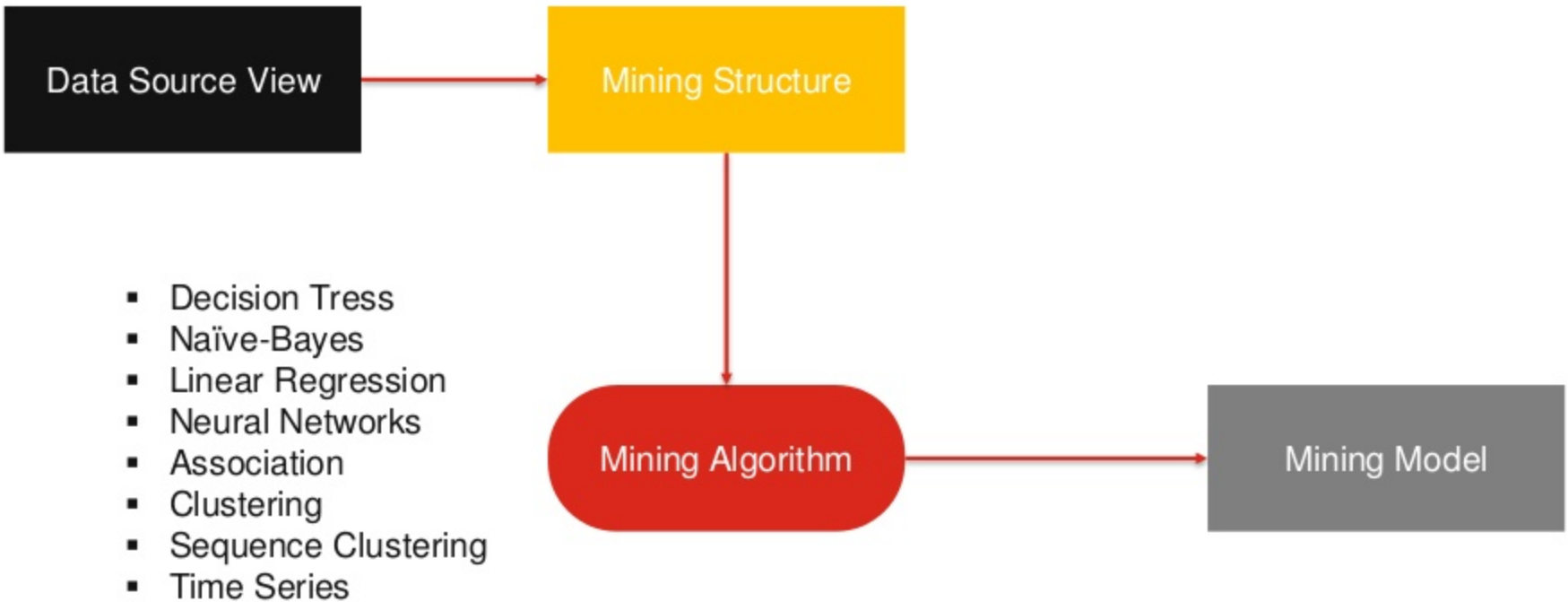


 |

Data Science and Machine Learning

Microsoft Analysis Services

Hitachi Consulting
We make it happen. Better.

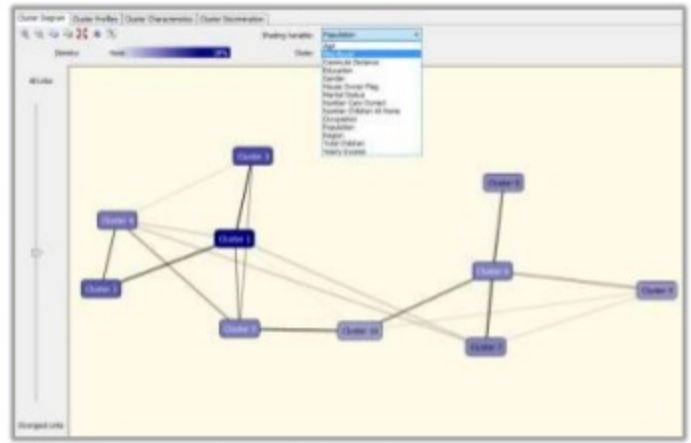


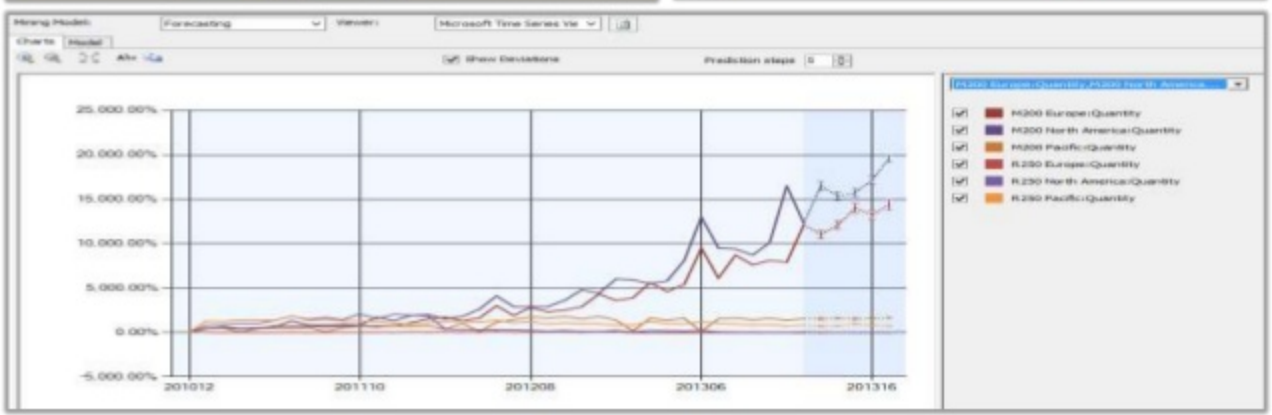
Explore/ Interpret Model

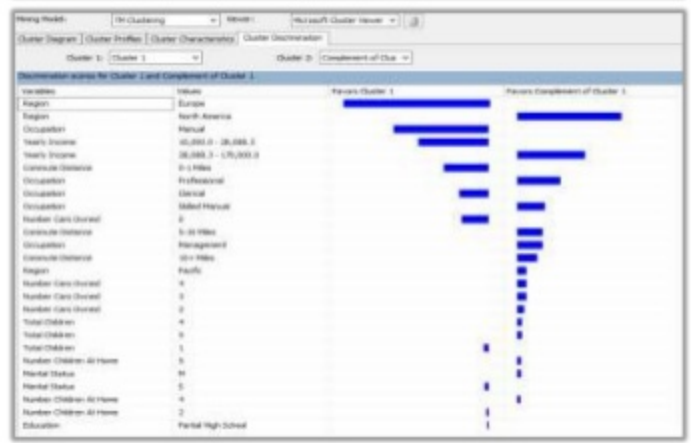
Data Science and Machine Learning

Microsoft Analysis Services

Hitachi Consulting
We make it happen. Better.







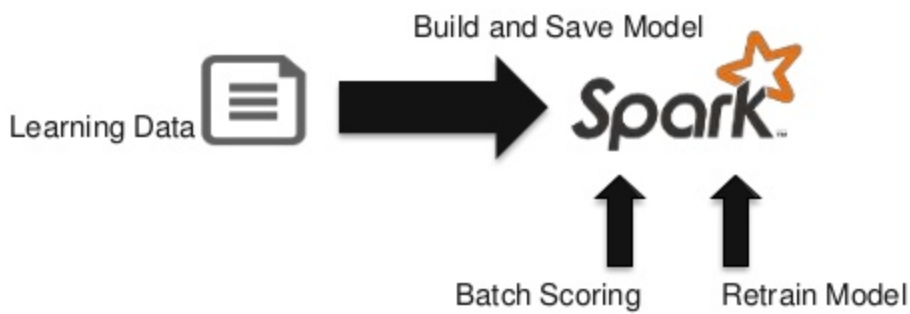
 |

Hitachi Consulting
We make it happen. Better.

Data Science and Machine Learning

Spark ML on HDInsight

- HDInsight-based ML Services provide by Spark
- Spark MLlib and Spark ML functionality
- Stream mining functionality
- Imports data from HDFS (Blob Storage/Data Lake)
- Scalable – Highly Distributed
- Extensible – Python and R
- Suitable for Big Data - **Batch Model Training and Data Scoring**



Data Science and Machine Learning

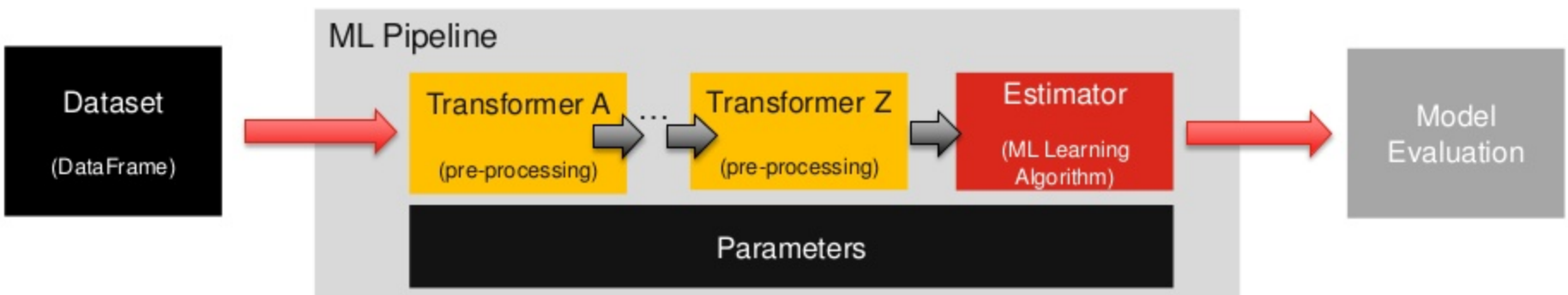
Spark ML on HDInsight – Pipelines

Hitachi Consulting
We make it happen. Better.

Spark

Spark ML standardizes APIs for machine learning algorithms to make it easier to combine multiple algorithms into a single pipeline, or workflow.

- **Transformers** – used for data pre-processing. Input: DataFrame - Output:DataFrame
- **Estimators** – ML algorithm used to build a predictive model. Input: DataFrame - Output: Model.
- **Parameters** – Configurations for Transformers and Estimators
- **Pipeline** – Chains Transformers and Estimators



 |

Data Science and Machine Learning

Spark ML on HDInsight - functionality

Hitachi Consulting
We make it happen. Better.

Spark

Transformers

Text Feature Extraction
- TF-IDF (HashingTF and IDF)
- Word2Vec
- CountVectorizer
- Tokenizer
- StopWordsRemover
- n-gram

Feature Selection
- VectorSlicer
- RFormula
- ChiSqSelector

Dimensionality Reduction
- PCA

Features Vector Preparation
- VectorAssembler
- VectorIndexer
- StringIndexer
- IndexToString

Feature Type Conversion
- Binarizer
- Discrete Cosine Transform (DCT)
- OneHotEncoder
- Bucketizer
- QuantileDiscretizer

Feature Scaling
- Normalizer
- StandardScaler
- MinMaxScaler

Feature Construction
- SQLTransformer
- ElementwiseProduct
- PolynomialExpansion

Estimators (supervised)

Classification
- Decision Trees – Ensembles
- Naïve-Bayes
- SVM

Regression
- Linear Regression
- SVM

Other (Unsupervised)

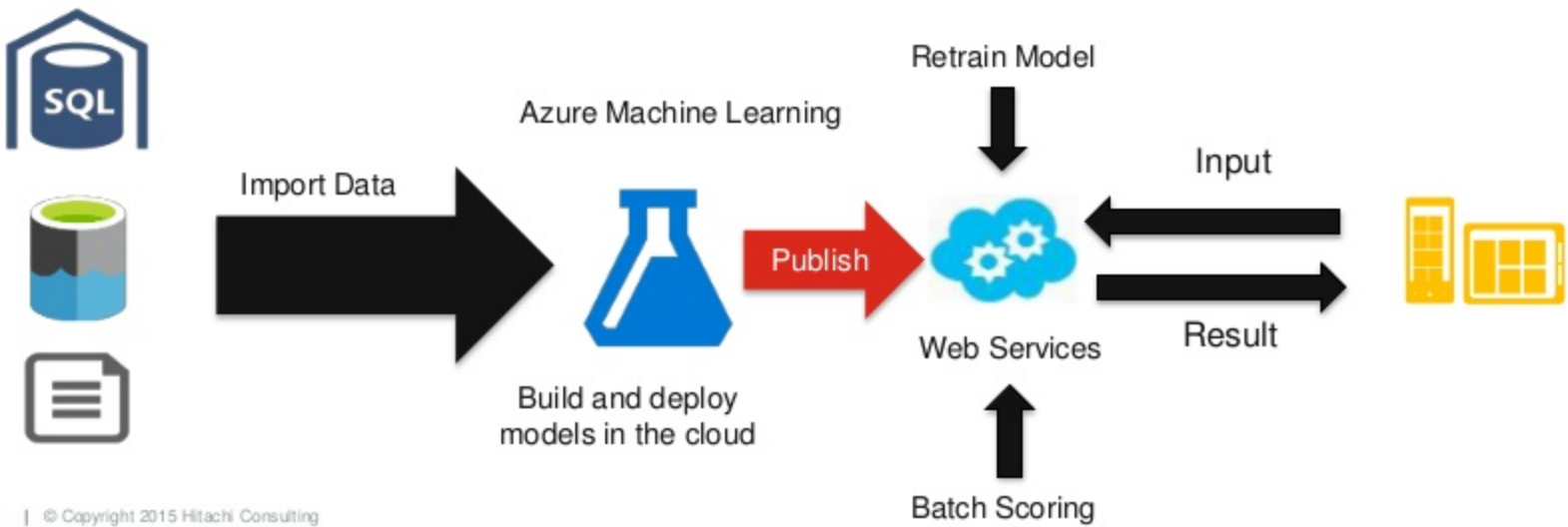
Clustering
Collaborative Filtering
Frequent Pattern Mining

Hitachi Consulting
We make it happen. Better.

Data Science and Machine Learning

Azure Machine Learning

- Cloud-based Machine Learning Services
- Imports data from everywhere
- Easy to **develop and productionize** – Web Services
- Extensible via R and Python scripts
- Limited in exploration and model interpretation
- Rich pre-processing functionality

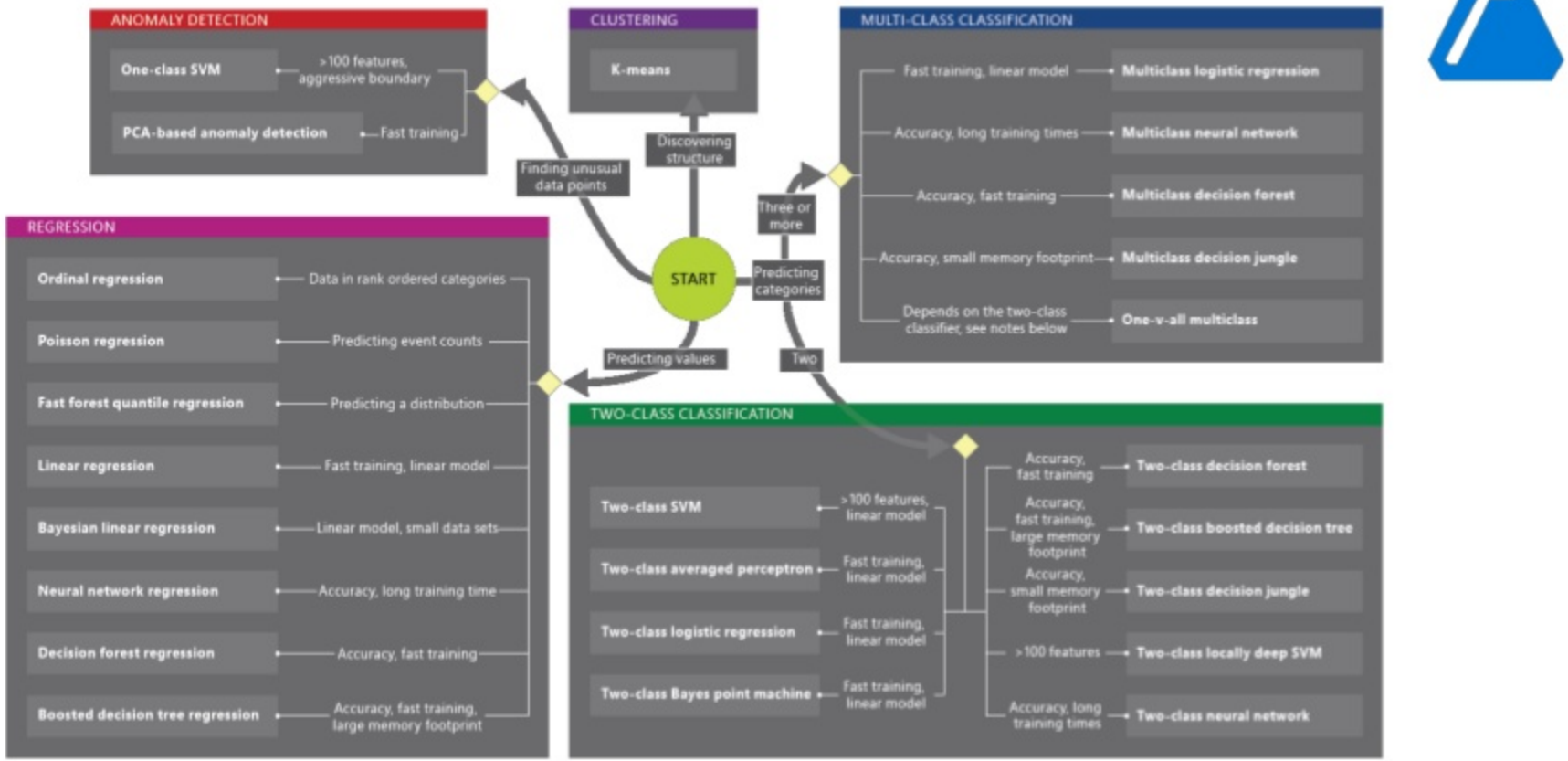


 |

Hitachi Consulting
We make it happen. Better.

Data Science and Machine Learning

Azure Machine Learning – algorithms cheat sheet

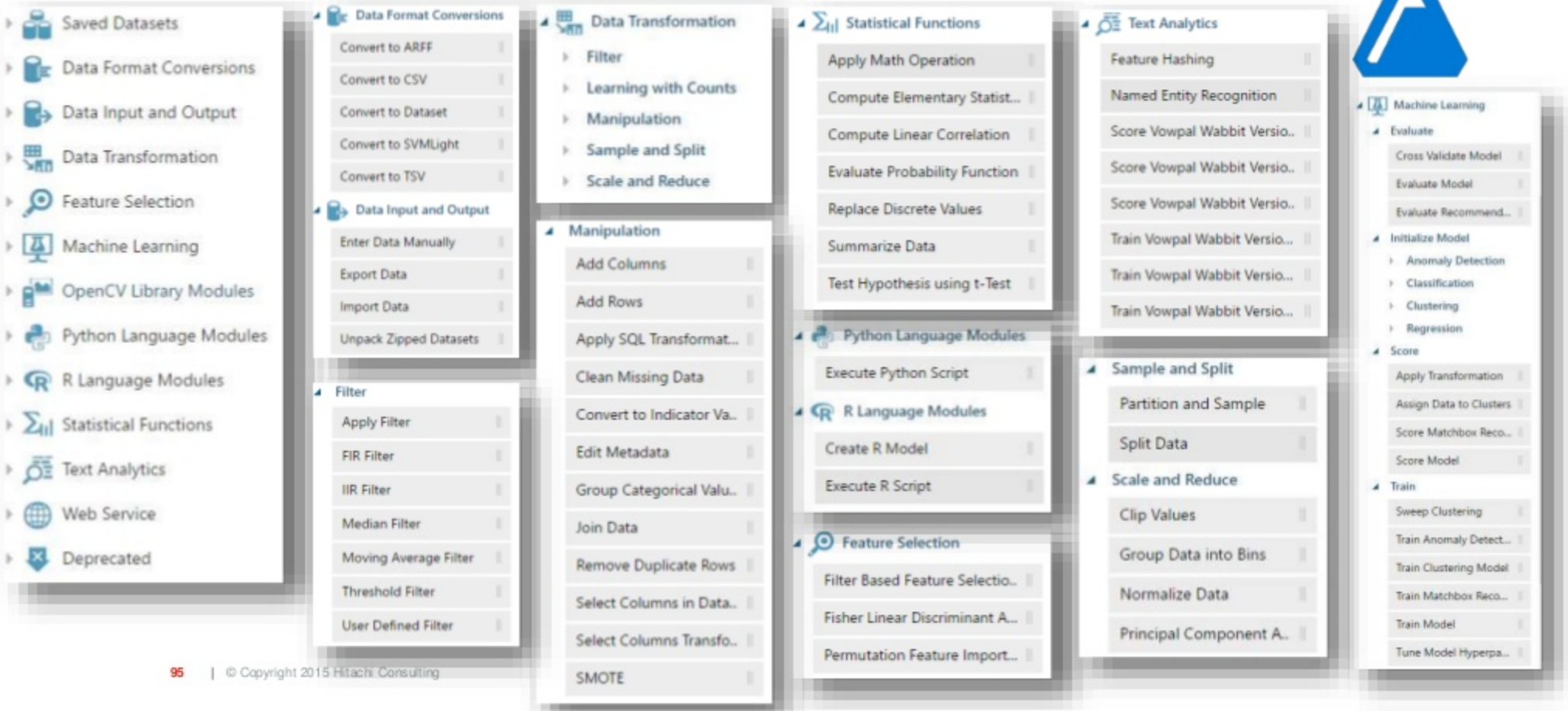


 |

Data Science and Machine Learning

Azure Machine Learning – ML Studio

Hitachi Consulting
We make it happen. Better.

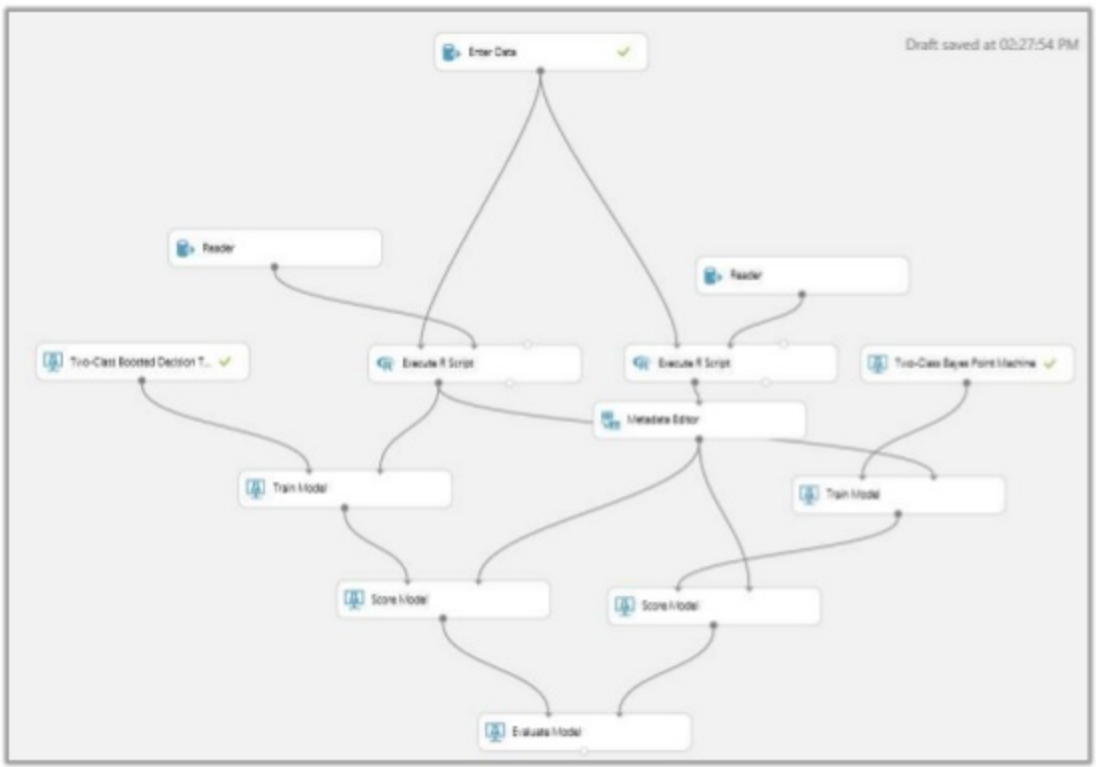


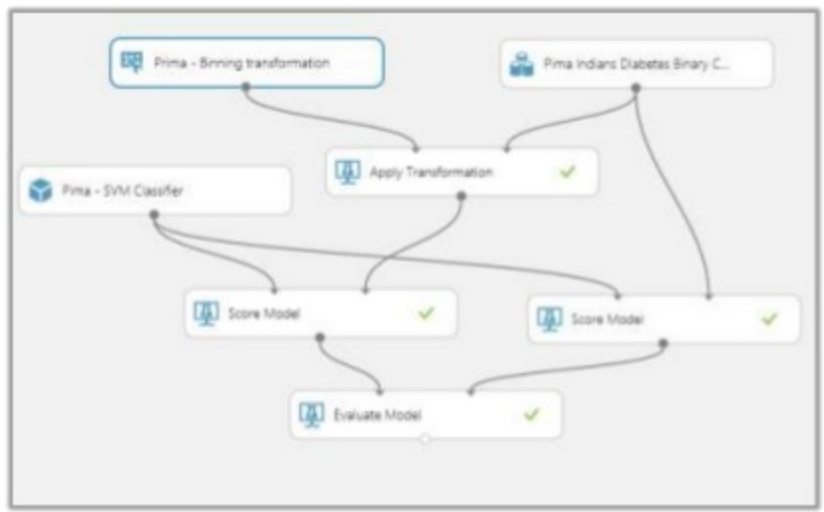
 |

Data Science and Machine Learning

Azure Machine Learning – ML Studio

Hitachi Consulting
We make it happen. Better.





Data Science and Machine Learning

Microsoft R Server

Microsoft R Open (MRO)

- Based on latest Open Source R (3.2.2.) - Built, tested, and distributed by Microsoft
- **Enhanced by Intel Math Kernel Library** (MKL) to speed up linear algebra functions
- More **efficient and multi-threaded** math computation
- Revolutions Open-Source R packages (**ParallelR, Rhadoop, AzureML**, etc.)
- Compatible with all R-related software

 |

Data Science and Machine Learning

Microsoft R Server

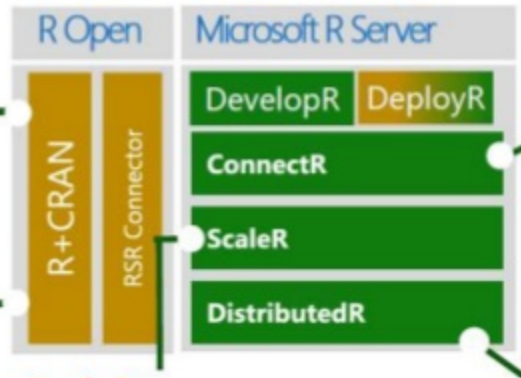
Hitachi Consulting
We make it happen. Better.

R+CRAN

- Open source R interpreter
  - R 3.1.2
- Freely-available huge range of R algorithms
- Algorithms callable by RevoR
- Embeddable in R scripts
- 100% Compatible with existing R scripts, functions and packages

MRO

- Performance enhanced R interpreter
- Based on open source R
- Adds high-performance math library to speed up linear algebra functions



ScaleR

- Ready-to-Use high-performance big data big analytics
- Fully-parallelized analytics
- Data prep & data distillation
- Descriptive statistics & statistical tests
- Range of predictive functions
- User tools for distributing customized R algorithms across nodes
- Wide data sets supported – thousands of variables

ConnectR

- High-speed & direct connectors

Available for:

- High-performance XDF
- SAS, SPSS, delimited & fixed format text data files
- Hadoop HDFS (text & XDF)
- Teradata Database & Aster
- EDWs and ADWs
- ODBC

DistributedR

- Distributed computing framework
- Delivers cross-platform portability

Hitachi Consulting
We make it happen. Better.

Data Science and Machine Learning

Microsoft R Server

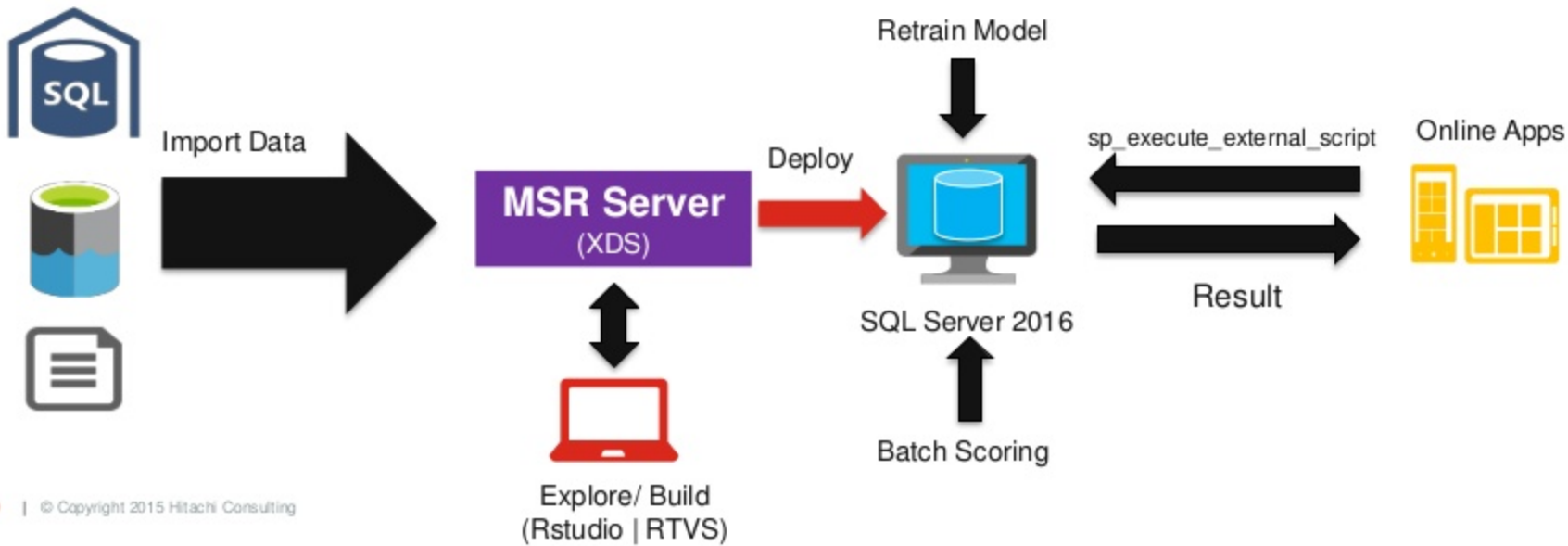
	CRAN	MRO	MRS
Data size	In-memory	In-memory	In-memory & disk
Efficiency	Single threaded	Multi-threaded	Multi-threaded, parallel processing 1:N servers
Support	Community	Community	Community + Commercial
Functionality	7500+ innovative analytic packages	7500+ innovative analytic packages	7500+ innovative packages + commercial parallel high-speed functions
Licence	Open Source	Open Source	Commercial license.

Hitachi Consulting
We make it happen. Better.

Data Science and Machine Learning

Microsoft R Server

- Powerful and scalable
- Installed on **Windows, Linux, or as R Services** in SQL Server 2016
- Work in different compute contexts – **local, Hadoop, Spark, or SQL Server 2016**
- Very suitable for interactive Data Science activities
- Models are deployed to **SQL Sever 2016, AzureML**, **or Web Service on R Server (DeployR)**



 |

Hitachi Consulting
We make it happen. Better.

Data Science and Machine Learning

Microsoft R Server – ScaleR functionality

Data Preparation

- Data import – Delimited, Fixed, SAS, SPSS, OBDC
- Variable creation & transformation
- Recode variables
- Factor variables
- Missing value handling
- Sort, Merge, Split
- Aggregate by category (means, sums)

Descriptive Statistics

- Min / Max, Mean, Median (approx.)
- Quantiles (approx.)
- Standard Deviation
- Variance
- Correlation
- Covariance
- Sum of Squares (cross product matrix for set variables)
- Pairwise Cross tabs
- Risk Ratio & Odds Ratio
- Cross-Tabulation of Data (standard tables & long form)
- Marginal Summaries of Cross Tabulations

Statistical Tests

- Chi Square Test
- Kendall Rank Correlation
- Fisher's Exact Test
- Student's t-Test

Sampling

- Subsample (observations & variables)
- Random Sampling

Predictive Models

- Sum of Squares (cross product matrix for set variables)
- Multiple Linear Regression
- Generalized Linear Models (GLM) exponential family distributions: binomial, Gaussian, inverse Gaussian, Poisson, Tweedie. Standard link functions: cauchit, identity, log, logit, probit. User defined distributions & link functions.
- Covariance & Correlation Matrices
- Logistic Regression
- Classification & Regression Trees
- Predictions/scoring for models
- Residuals for all models

Variable Selection

- Stepwise Regression

Simulation

- Simulation (e.g. Monte Carlo)
- Parallel Random Number Generation

Cluster Analysis

- K-Means

Classification

- Decision Trees
- Decision Forests
- Gradient Boosted Decision Trees
- Naïve Bayes

Combination

- rxDataStep
- rxExec
- PEMA-R API Custom Algorithms

 |

Data Science and Machine Learning

Azure Cognitive Services - SaaS

Hitachi Consulting
We make it happen. Better.

Language

Allow your apps to process natural language, evaluate sentiment and topics, and learn how to recognise what users want.

Language Understanding Intelligent Service

Teach your apps to understand commands from your users

Text Analytics API

Easily evaluate sentiment and topics to understand what users want

Web Language Model API

Use the power of predictive language models trained on web-scale data

Bing Spell Check API

Detecting and correcting spelling mistakes in your app

Speech

Processing spoken language in your applications

Bing Speech API

Convert speech to text and back again to understand user intent

Speaker Recognition API

Use speech to identify and authenticate individual speakers

Vision

State-of-the-art image processing algorithms to build more personalised apps by returning smart insights such as faces, images and emotion recognition.

Face API

Detect, analyse, organise and tag faces in photos

Emotion API

Personalise user experiences with emotion recognition

Search

Make your apps, web pages and other experiences smarter and more engaging with the Bing Search APIs.

Bing Search APIs

Search, image, video and news APIs for your apps

Bing Autosuggest API

Give your app intelligent autosuggest options for searches

Knowledge

Map complex information and data in order to solve tasks such as intelligent recommendations and semantic search.

Recommendations API

Predict and recommend items that your customers want

Azure Data Catalog

 |

Hitachi Consulting
We make it happen. Better.

Azure Data Catalog

Enterprise-wide metadata catalogue for data assets discovery

A cloud-based service into which data source can be registered, described and discovered by enterprise-wide users

- Registration of central data sources
- Self-service business intelligence
- Capturing tribal knowledge

Register
- Data Sources
- Data Assets
- Locatoins

Discover
- Search
- Browse
- Filter

Understand
- Metadata
- Expert
- Description

Consume
- Your data
- Your Tools
- Your Way

Contribute
- Tag
- Document
- Publish

Azure Data Catalog

Terminology

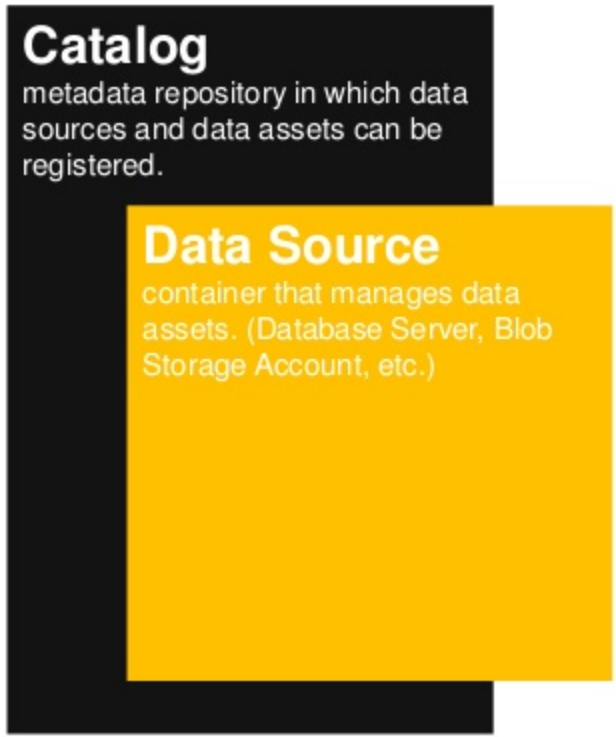
Hitachi Consulting
We make it happen. Better.

Catalog
metadata repository in which data sources and data assets can be registered.

Hitachi Consulting
We make it happen. Better.

Azure Data Catalog

Terminology



Hitachi Consulting
We make it happen. Better.

Azure Data Catalog

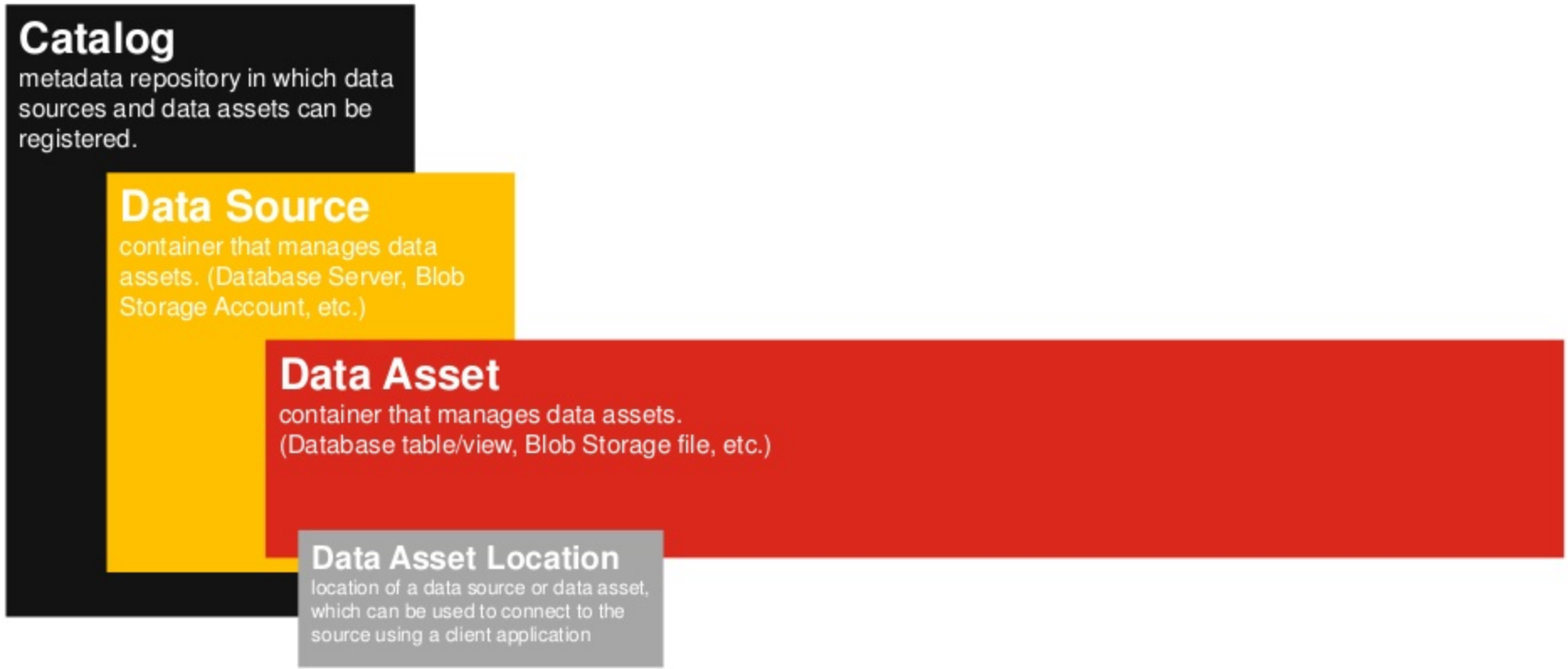
Terminology

 |

Hitachi Consulting
We make it happen. Better.

Azure Data Catalog

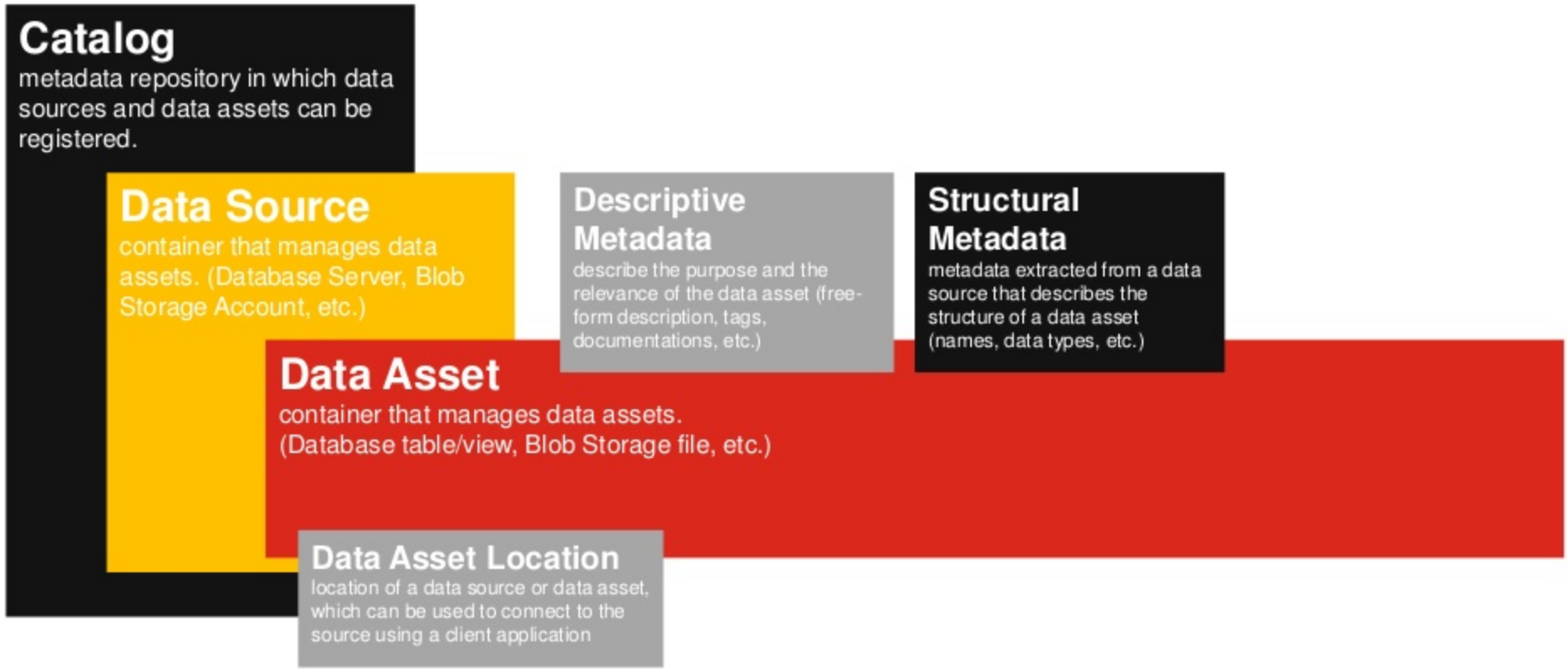
Terminology



Hitachi Consulting
We make it happen. Better.

Azure Data Catalog

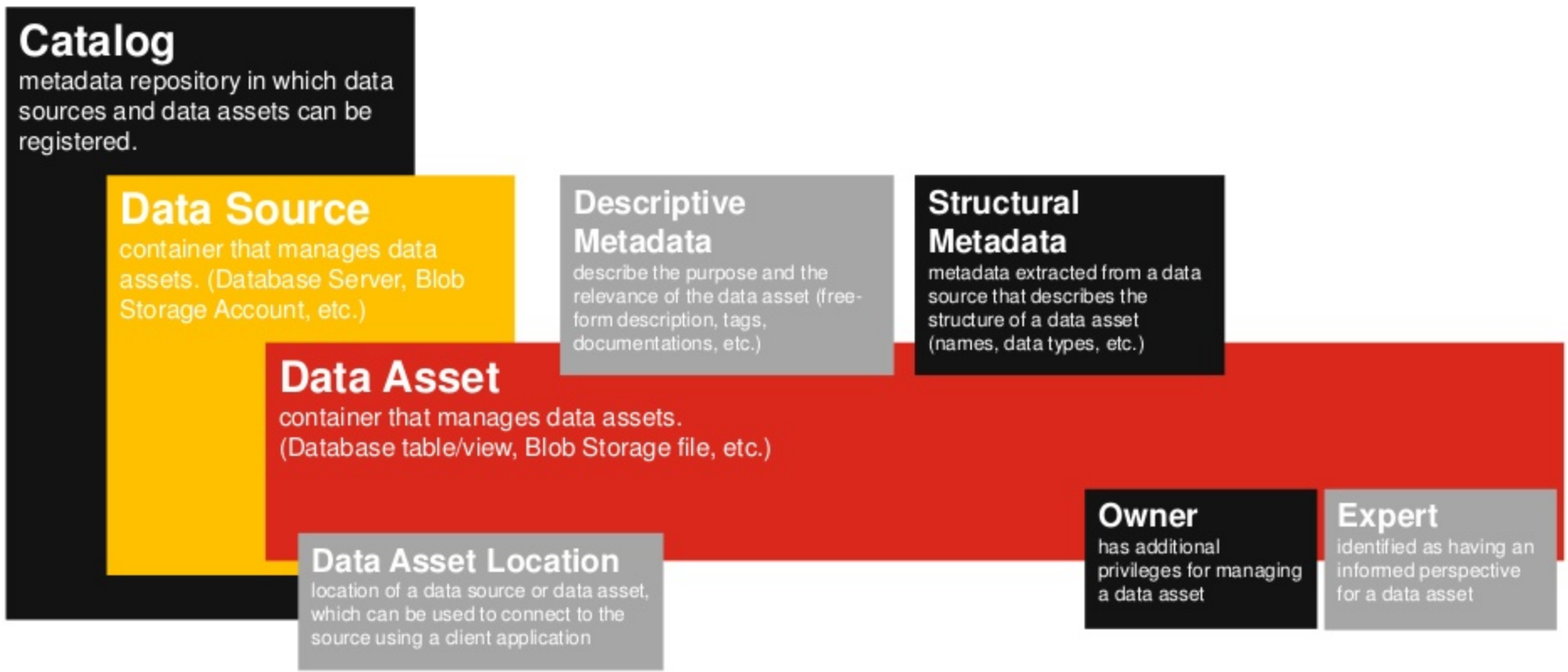
Terminology



Hitachi Consulting
We make it happen. Better.

Azure Data Catalog

Terminology



Azure Data Catalog

Terminology

Hitachi Consulting
We make it happen. Better.

Catalog
metadata repository in which data sources and data assets can be registered.

Data Source
container that manages data assets. (Database Server, Blob Storage Account, etc.)

Descriptive Metadata
describe the purpose and the relevance of the data asset (free-form description, tags, documentations, etc.)

Structural Metadata
metadata extracted from a data source that describes the structure of a data asset (names, data types, etc.)

Preview
a snapshot of up to 20 records that can be extracted from the data source during registration, and stored in the catalog

Data Asset
container that manages data assets.
(Database table/view, Blob Storage file, etc.)

Owner
has additional privileges for managing a data asset

Expert
identified as having an informed perspective for a data asset

Data Asset Location
location of a data source or data asset, which can be used to connect to the source using a client application

Profile
a snapshot of table-level and column-level metadata about the content of a registered data asset

Hitachi Consulting
We make it happen. Better.

Azure Data Catalog

Terminology

Glossary
Terms, parent terms, and definitions

Catalog
metadata repository in which data sources and data assets can be registered.

Data Source
container that manages data assets. (Database Server, Blob Storage Account, etc.)

Descriptive Metadata
describe the purpose and the relevance of the data asset (free-form description, tags, documentations, etc.)

Structural Metadata
metadata extracted from a data source that describes the structure of a data asset (names, data types, etc.)

Preview
a snapshot of up to 20 records that can be extracted from the data source during registration, and stored in the catalog

Data Asset
container that manages data assets. (Database table/view, Blob Storage file, etc.)

Owner
has additional privileges for managing a data asset

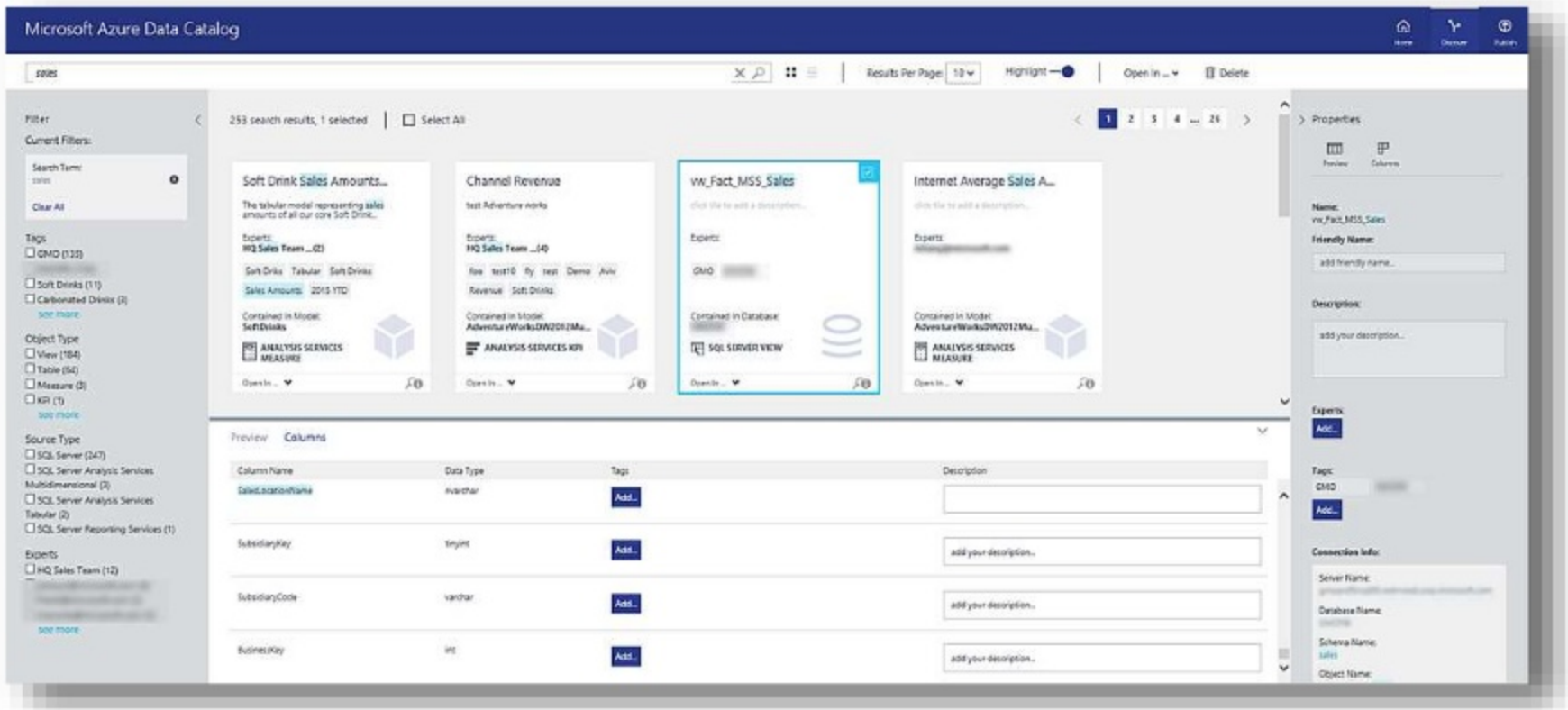
Expert
identified as having an informed perspective for a data asset

Data Asset Location
location of a data source or data asset, which can be used to connect to the source using a client application

Profile
a snapshot of table-level and column-level metadata about the content of a registered data asset

Azure Data Catalog

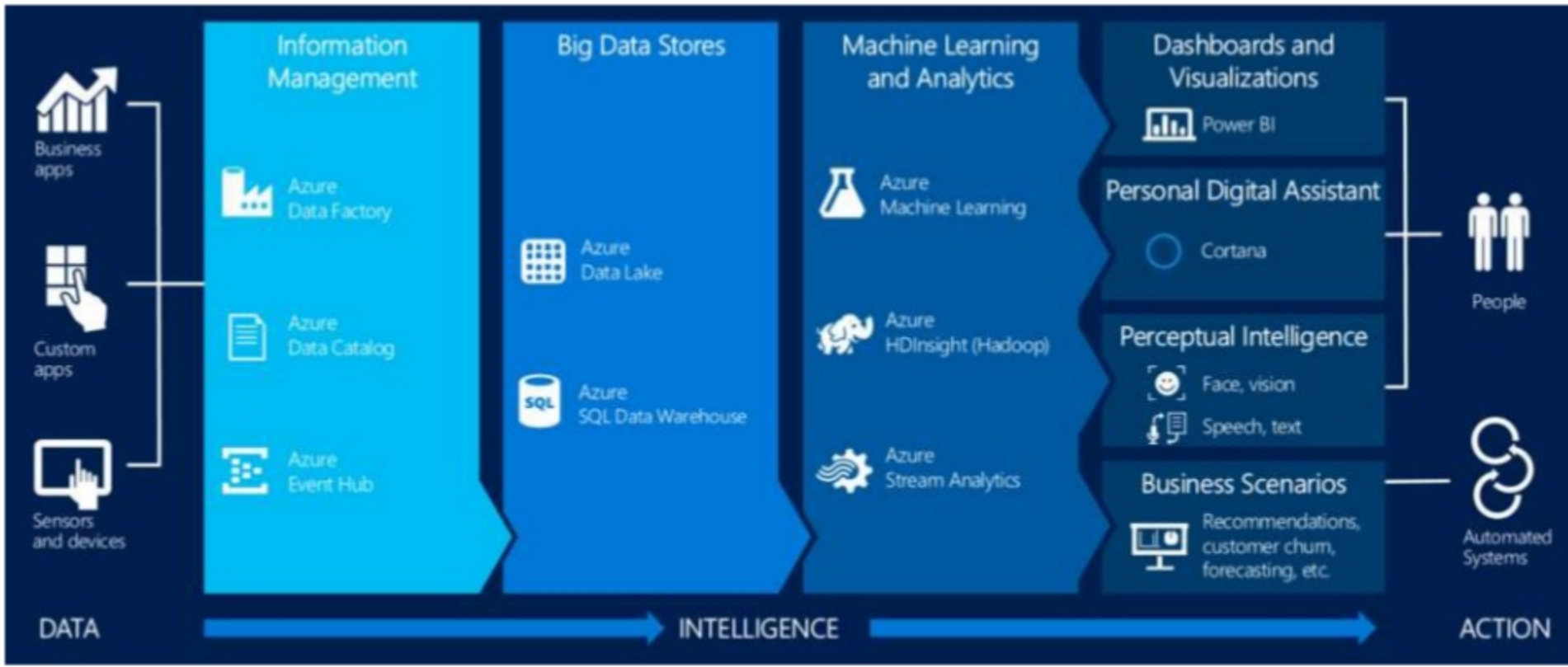
Preview



 |

Cortana Analytical Suite

Hitachi Consulting
We make it happen. Better.



https://gallery.cortanaintelligence.com

 |

Hitachi Consulting
We make it happen. Better.

Acknowledgement

This guy is awesome

Thanks to Paul Lineham, the best big data architect I worked with, for supplying me with practical ideas and knowledge on architecting and delivering data platforms.

Hitachi Consulting
We make it happen. Better.

My Background

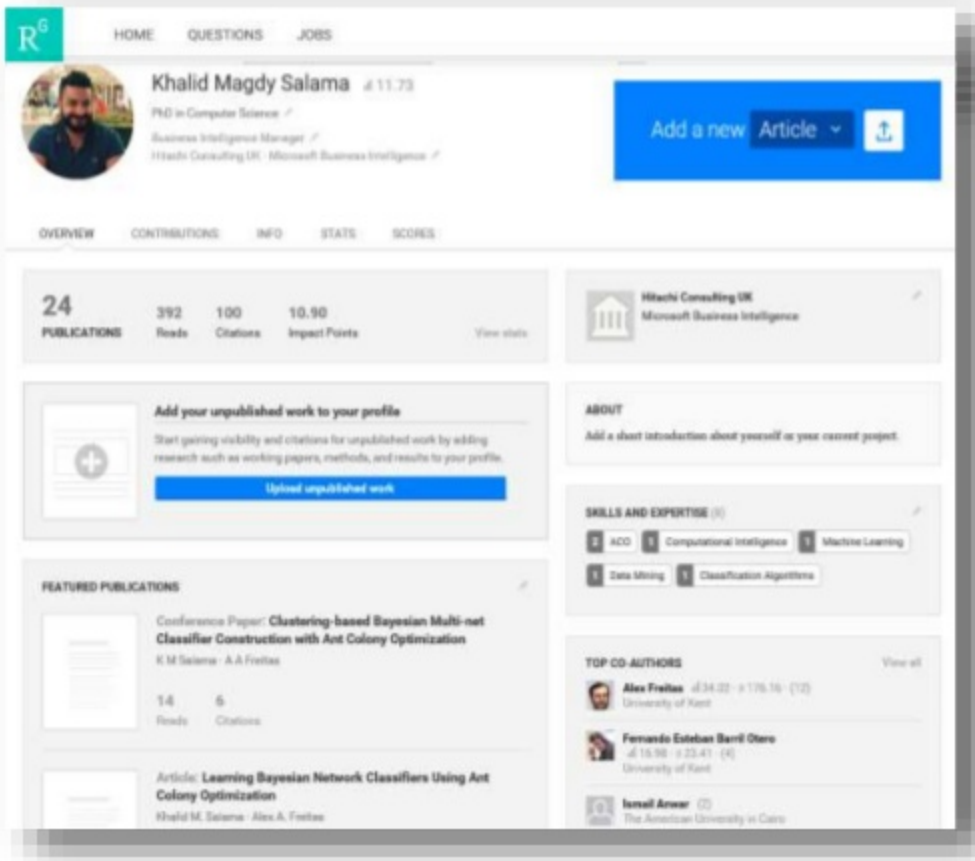
Applying Computational Intelligence in Data Mining

- Honorary Research Fellow, School of Computing , University of Kent.
- Ph.D. Computer Science, University of Kent, Canterbury, UK.
- M.Sc. Computer Science , The American University in Cairo, Egypt.

- 25+ published journal and conference papers, focusing on:
  - *classification rules induction,*
  - *decision trees construction,*
  - *Bayesian classification modelling,*
  - *data reduction,*
  - *instance-based learning,*
  - *evolving neural networks*, and
  - *data clustering*

- **Journals**: *Swarm Intelligence, Swarm & Evolutionary Computation, , Applied Soft Computing, and Memetic Computing.*

- **Conferences**: *ANTS, IEEE CEC, IEEE SIS, EvoBio, ECTA, IEEE WCCI and INNS-BigData.*



ResearchGate.org

 |

Thank you!

 |