Azure HDInsight

Alex Zhao
Oct 2018

Microsoft

What is HDInsight?

Azure HDInsight is a cloud distribution of the Hadoop components from the Hortonworks Data Platform (HDP). Apache Hadoop was the original open-source framework for distributed processing and analysis of big data sets on clusters of computers.

The Hadoop technology stack includes related software and utilities, including Apache Hive, HBase, Spark, Kafka, and many others. To see available Hadoop technology stack components on HDInsight.

Why Hadoop in the Cloud?

No HW costs

Unlimited scale

Pay what you need

Deployed in minutes

Hadoop is also a cluster type that has:

- YARN for job scheduling and resource management
- MapReduce for parallel processing
- The Hadoop distributed file system (HDFS)

Hadoop clusters are most often used for batch processing of stored data. Other kinds of clusters in HDInsight have additional capabilities

No hardware costs

Hadoop in the Cloud bypasses hardware costs

Hardware acquisition

Hardware maintenance

Performance tuning

Unlimited Scale

Hadoop in the Cloud bypasses capacity planning

Spin up any number of Hadoop nodes on-demand

Go from tens of nodes to thousands of nodes

Pay for What You Need

Hadoop is billed by usage

Billed for usage

Clusters can be deleted when no longer used

Pay what
you need

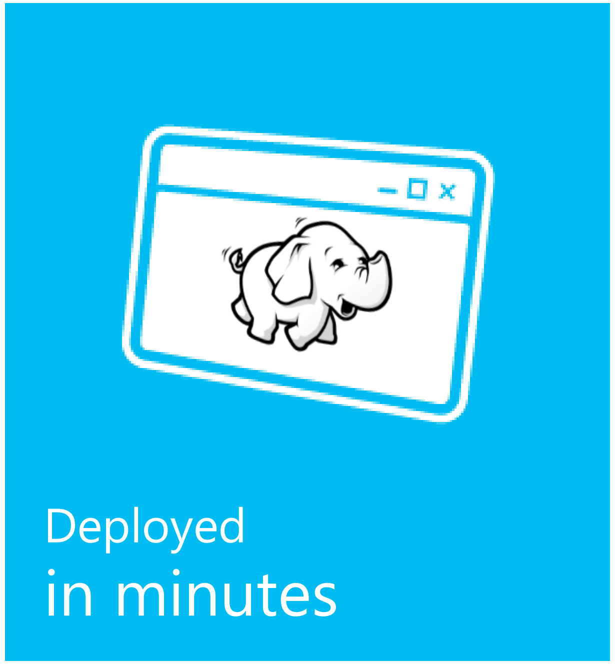
Deployed in minutes

Hadoop in the Cloud Bypasses deployment expertise

- Hadoop is non-trivial to install and get up and running on multi-nodes
- Education gap in IT community regarding Hadoop

Hadoop is deployed in minutes

- Spin up any number of Hadoop nodes on-demand
- Up and running in a few clicks (and within minutes)



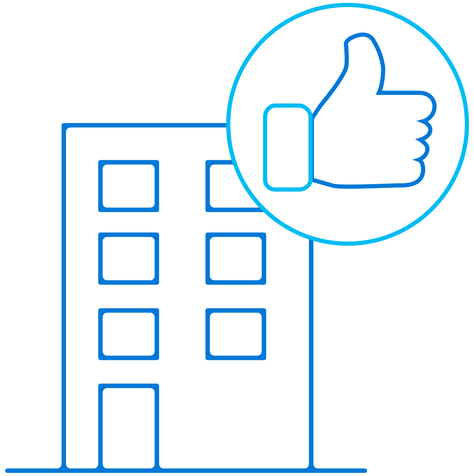
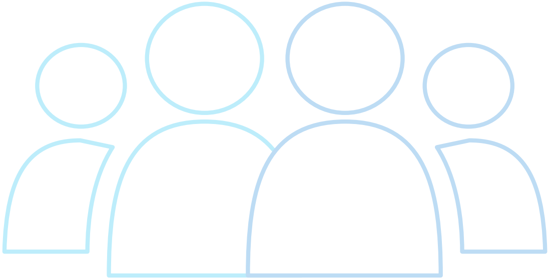
Azure HDInsight

Big Data made easy

Enterprise Ready	Easier and more productive for all users	Hybrid

Azure HDInsight

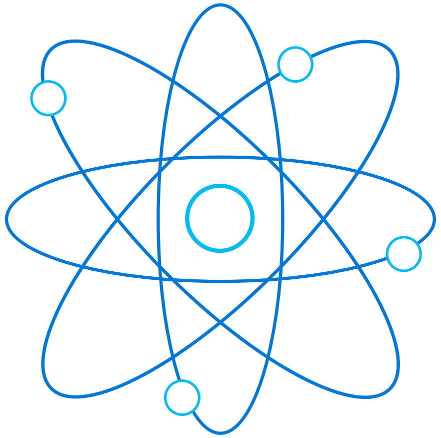
Big Data made easy

Enterprise Ready	Easier and more productive for all users	Hybrid
		

Highly Available – Designed for the cloud ground up

- HDInsight provides primary and secondary headnodes allowing for better reliability
- Have invested in making entire stack including Resource Manager, HiverServer2 HA ready
- HDInsight stack includes Zookeeper nodes at no extra charge to customer

Highest availability guarantee in the industry for peace of mind

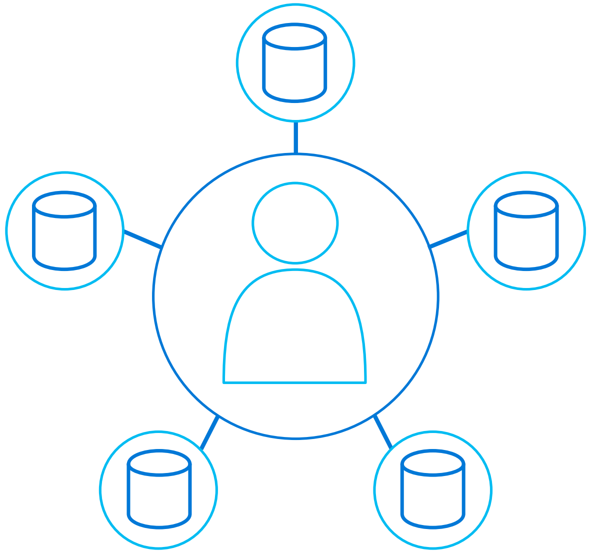


99.9% SLA

- Managed, monitored and supported by Microsoft
- Enterprise-leading SLA—99.9% uptime for both VM connectivity and Hadoop running in VMs
- No IT resources needed for upgrades and patching
- Microsoft monitors your deployment so you don’t have to

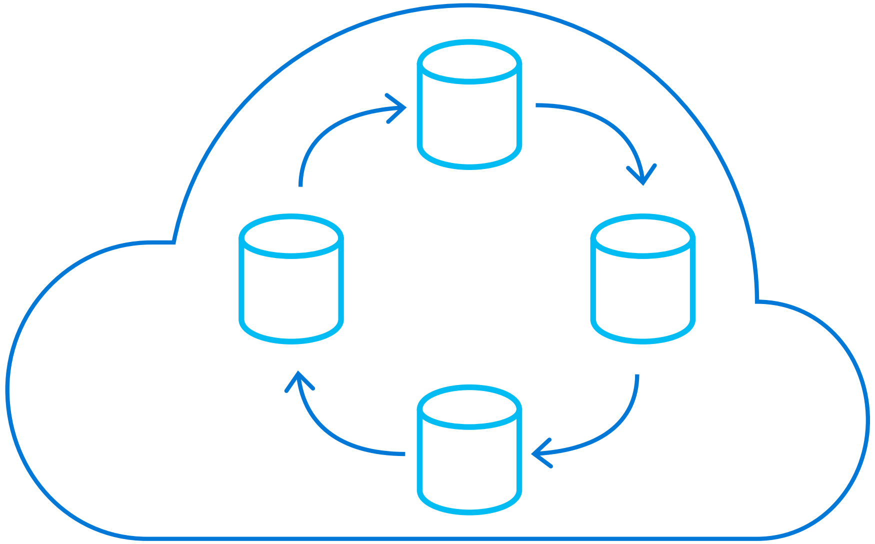
*Applies to HDInsight only

Always encrypted, Role-based security & Auditing



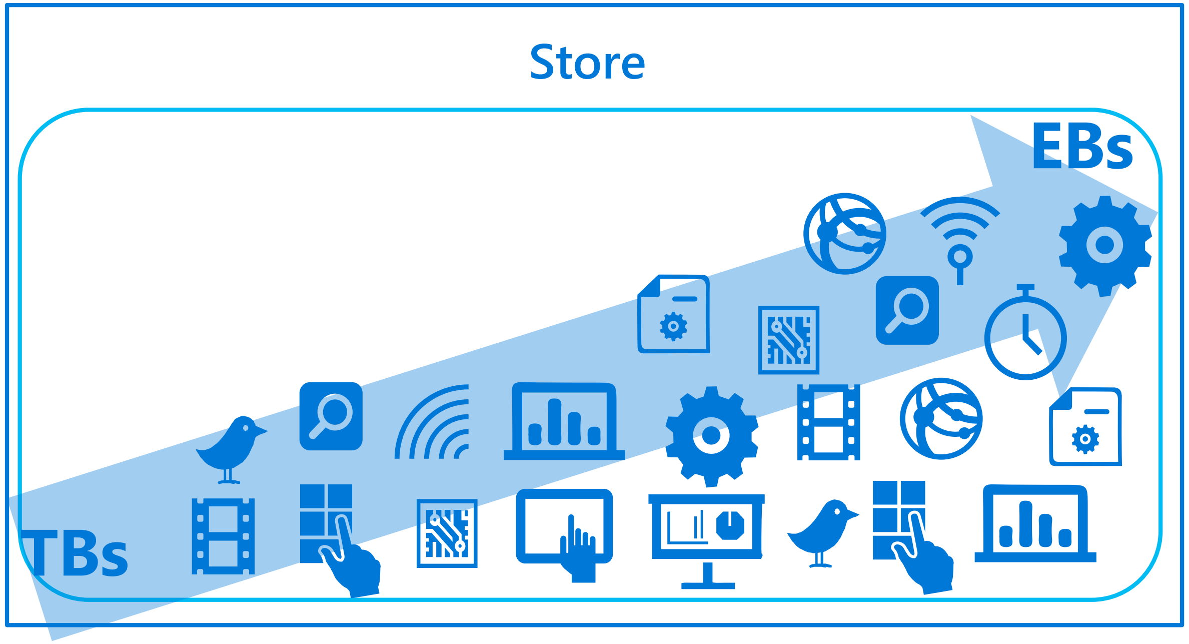
- Always encrypted; in motion using SSL, and at rest using keys in Azure Key Vault
- Single sign-on, multi-factor authentication and integration of on-premises identities w/Active Directory integration
- Fine-grained ACLs for role-based access controls with Apache Ranger
- Auditing every access / configuration change with Apache Ranger

Alerting, monitoring, and pre-emptive actions



- Enhanced workload protection through integration with Microsoft Operations Management Suite (OMS)
- Threat detection, monitoring, and management

Petabyte size files and Trillions of objects



- Store data in it’s native format
- PB sized files, 200x larger than anyone else
- Scalable throughput for massively parallel analytics
- No need to redesign application or reparation data at higher scale

Backed by Microsoft and Hortonworks

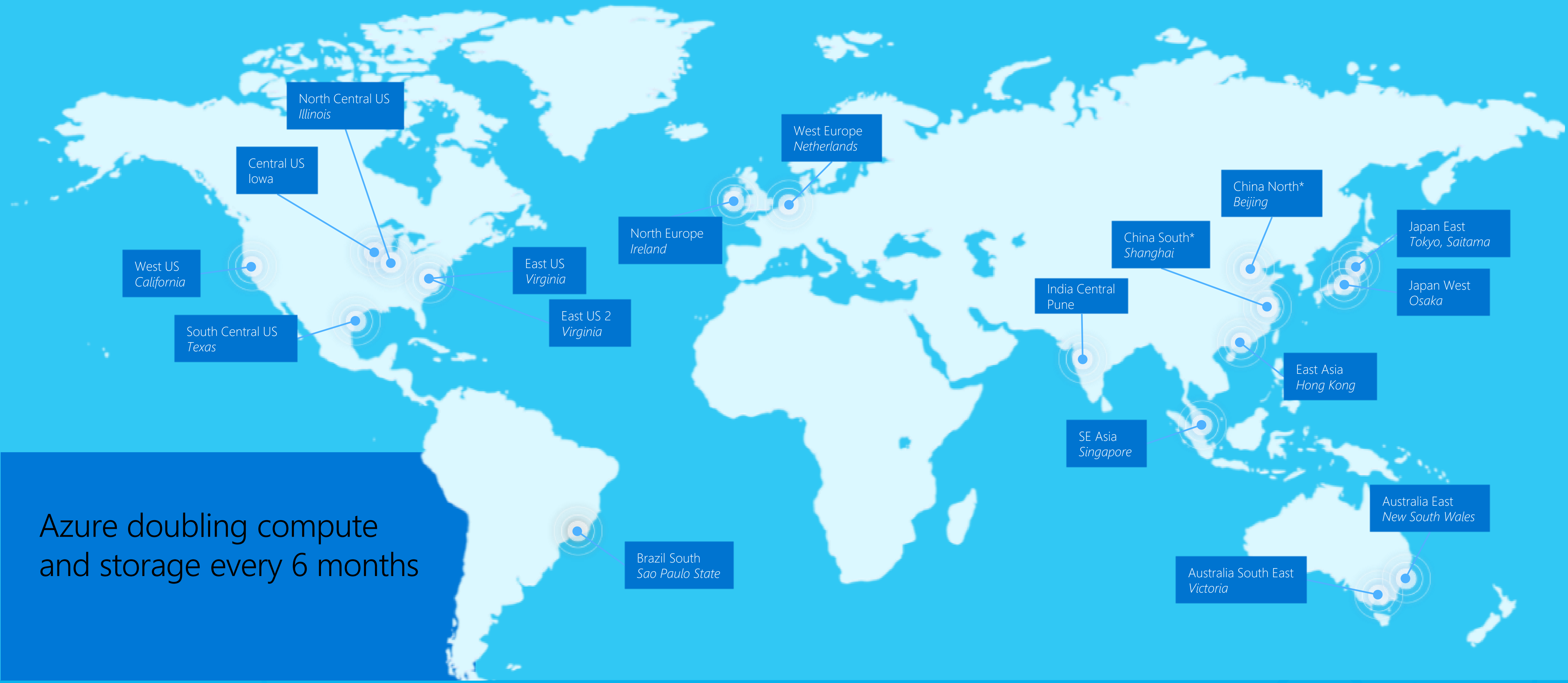
Microsoft

+

HORTONWORKS®
POWERING THE FUTURE OF DATA™

- Microsoft + Hortonworks has **37 committers** for Hadoop Core; more than all managed cloud Hadoop vendors combined
- Uniquely ready to support your deployment
- Can fix and commit code back to Hadoop

Runs in the most datacenters worldwide



Azure doubling compute and storage every 6 months

Lower total cost of ownership

- No hardware
- Hadoop support included with Azure support
- Pay only for what you use
- Independently scale storage and compute
- No need to hire specialized operations team
- 63% lower total cost of ownership than on-premises*

*IDC study "The Business Value and TCO Advantage of Apache Hadoop in the Cloud with Microsoft Azure HDInsight"

Azure HDInsight

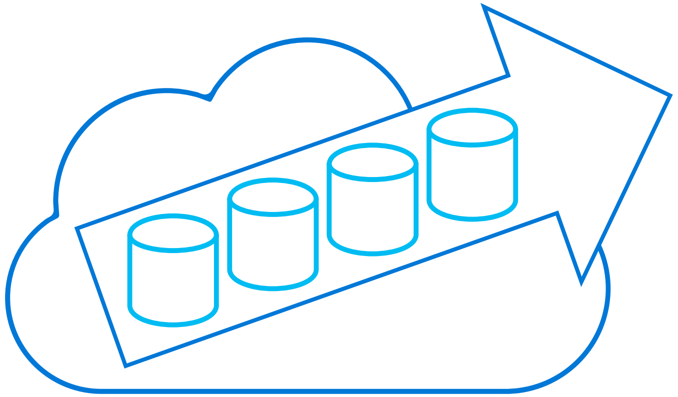
Big Data made easy

Enterprise Ready

Easier and more productive for all users

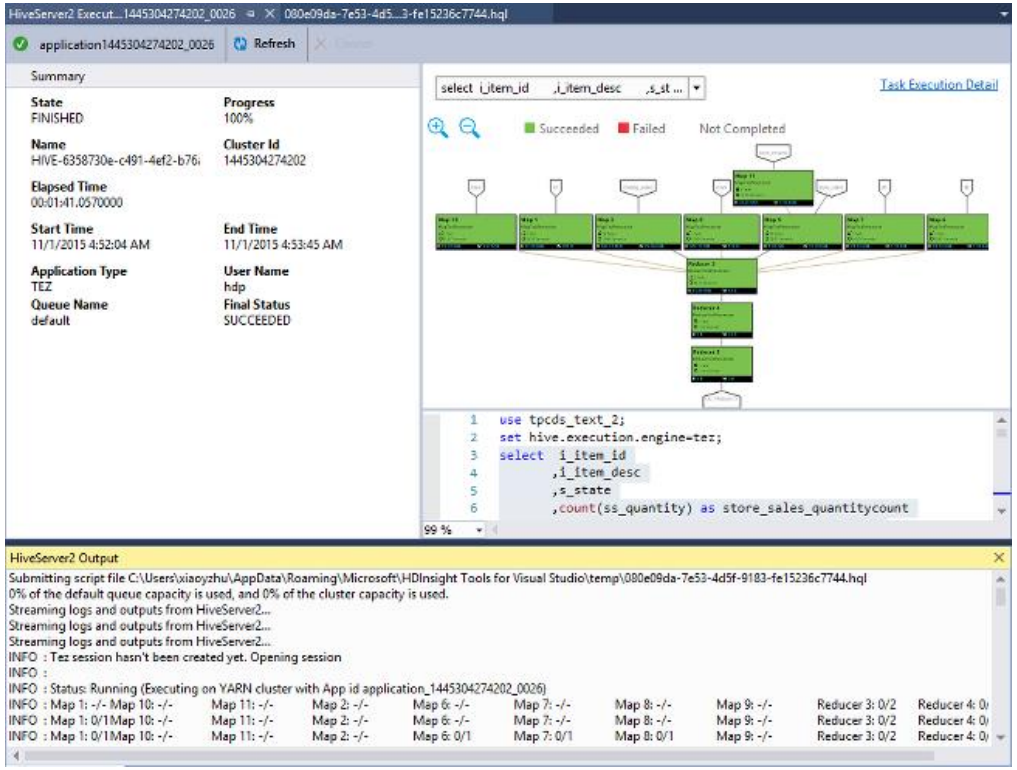
Hybrid

Easy for administrators to spin up quickly



- Deploy big data projects in minutes
- No hardware to install, tune, configure or deploy
- No infrastructure or software to manage
- Scale to tens to thousands of machines instantly

Debug and Optimize your Big Data programs with ease



HiveServer2 Execut...1445304274202_0026
080e09da-7e53-4d5...3-fe15236c7744.hql
application1445304274202_0026
Refresh
Summary
select i_item_id ,i_item_desc ,s_st ...
Task Execution Detail
🟩 Succeeded
🟥 Failed
Not Completed
99 %

State
FINISHED

Progress
100%

Name
HIVE-6358730e-c491-4ef2-b76i

Cluster Id
1445304274202

Elapsed Time
00:01:41.0570000

Start Time
11/1/2015 4:52:04 AM

End Time
11/1/2015 4:53:45 AM

Application Type
TEZ

User Name
hdp

Queue Name
default

Final Status
SUCCEEDED

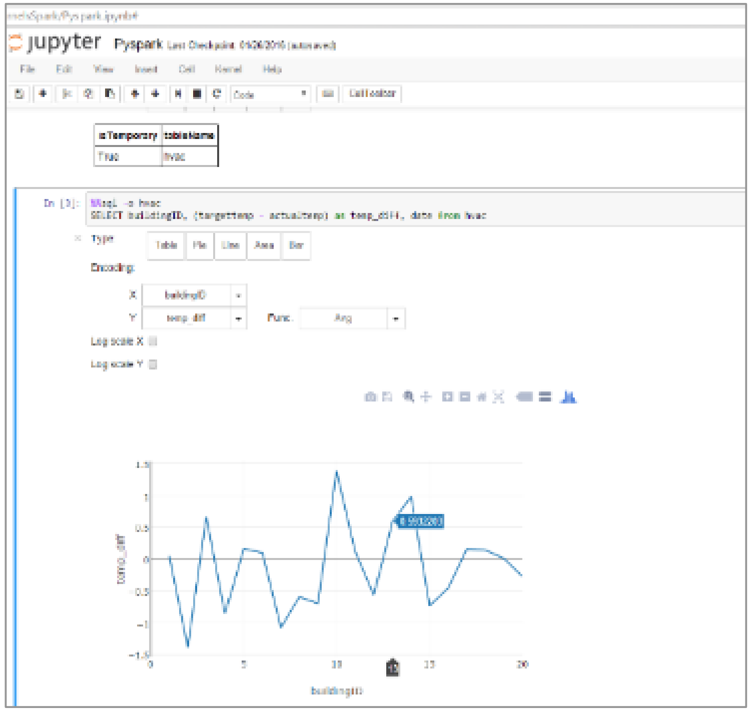
```
use tpcds_text_2;
set hive.execution.engine=tez;
select  i_item_id
       ,i_item_desc
       ,s_state
       ,count(ss_quantity) as store_sales_quantitycount
```

HiveServer2 Output

```
Submitting script file C:\Users\xiaoyzhu\AppData\Roaming\Microsoft\HDInsight Tools for Visual Studio\temp\080e09da-7e53-4d5f-9183-fe15236c7744.hql
0% of the default queue capacity is used, and 0% of the cluster capacity is used.
Streaming logs and outputs from HiveServer2...
Streaming logs and outputs from HiveServer2...
Streaming logs and outputs from HiveServer2...
INFO  : Tez session hasn't been created yet. Opening session
INFO  :
INFO  : Status: Running (Executing on YARN cluster with App id application_1445304274202_0026)
INFO  : Map 1: -/- Map 10: -/-     Map 11: -/-     Map 2: -/-     Map 6: -/-     Map 7: -/-     Map 8: -/-     Map 9: -/-     Reducer 3: 0/2     Reducer 4: 0/
INFO  : Map 1: 0/1Map 10: -/-      Map 11: -/-     Map 2: -/-     Map 6: -/-     Map 7: -/-     Map 8: -/-     Map 9: -/-     Reducer 3: 0/2     Reducer 4: 0/
INFO  : Map 1: 0/1Map 10: -/-      Map 11: -/-     Map 2: -/-     Map 6: 0/1     Map 7: 0/1     Map 8: 0/1     Map 9: -/-     Reducer 3: 0/2     Reducer 4: 0/
```

- Deep integration with IDEs for developer productivity: Visual Studio, Eclipse, & IntelliJ
- Integrated with Hive, Pig, Storm, and Spark
- Visually see execution of Hive jobs ran by the Tez execution engine
- Full Intellisense

Easy notebook experience for data engineers

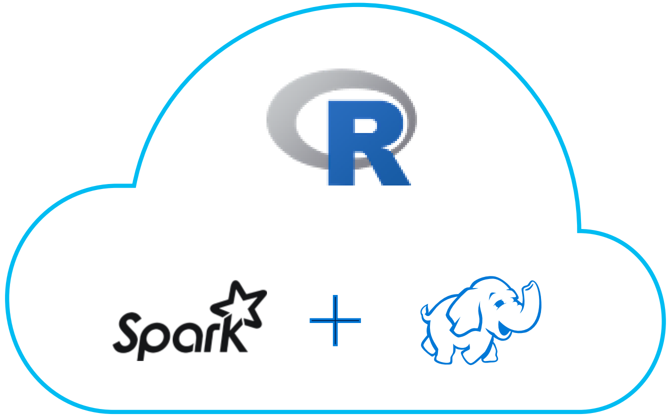


isTemporary	tableName
True	hvac

```
In [3]: %%sql -o hvac
        SELECT buildingID, (targettemp - actualtemp) as temp_diff, date from hvac
```

- Most popular notebooks, Jupyter and Zeppelin out-of-the-box
- Combine code, statistical equations and visualizations
- Worked w/ Jupyter community to enhance kernel to allow Spark execution through REST endpoint

Easy for data scientists with familiar R language

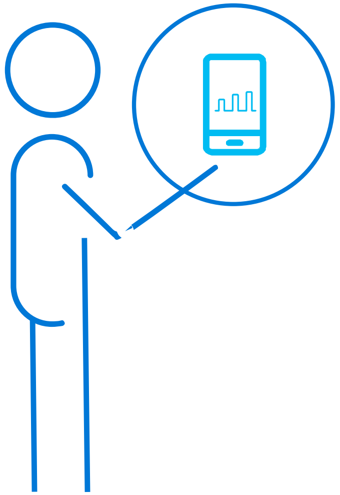


R Server for HDInsight

- Largest portable R parallel analytics library
- Terabyte-scale machine learning—1,000x larger than in open source R
- Up to 100x faster performance using Spark and optimized vector/math libraries
- Enterprise-grade security and support

*Applies to HDInsight only

Easy for business analysts with interactive reports over big data



- Interactive BI with big data
- Spark 2.0 integration
- Interactive Hive with LLAP- keeps data compressed running in-memory 25x faster
- ODBC driver to use Power BI or third party tools (Tableau, SAP, Qlik, etc.)

Azure HDInsight

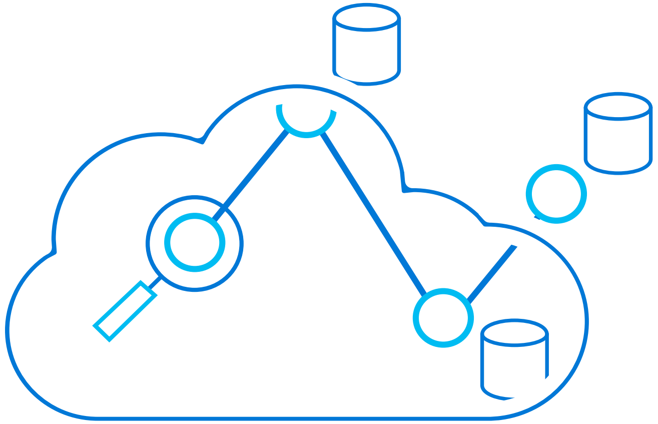
Big Data made easy

Enterprise Ready

Easier and more productive for all users

Hybrid

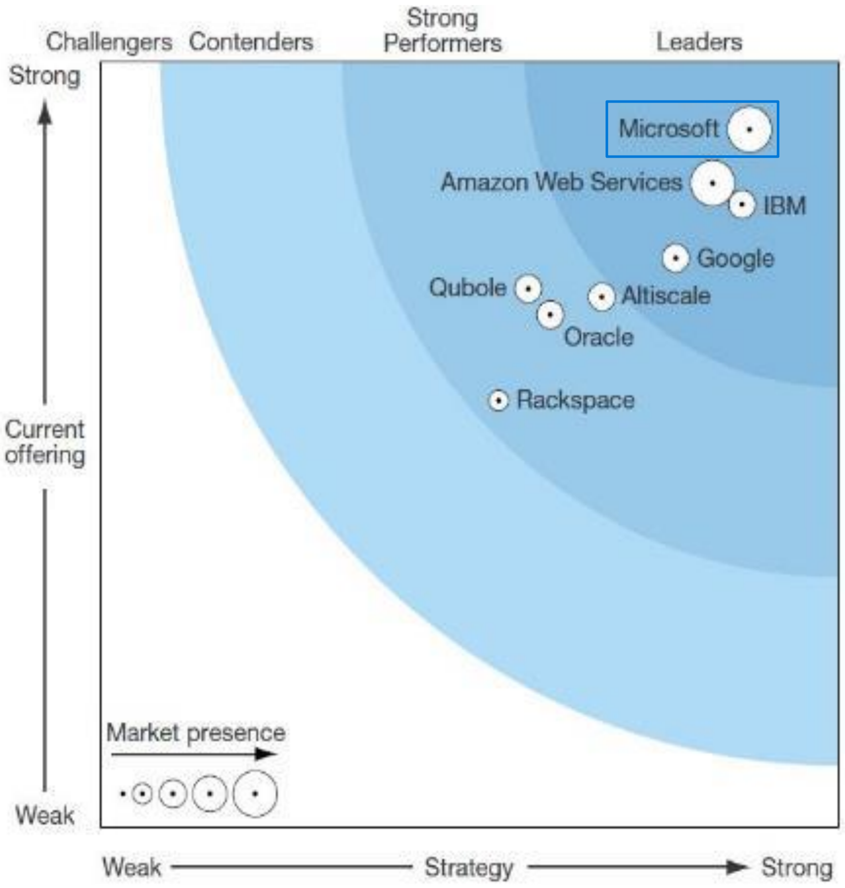
On-premises and cloud



- Uses Hortonworks Data Platform (HDP)
- Move projects from on-premises to cloud without code rewrite
- Hybrid scenarios supported like Dev/Test, burst, back up, disaster recovery

Recognized by top analysts

FORRESTER®



Forrester Wave for Big Data Hadoop Cloud

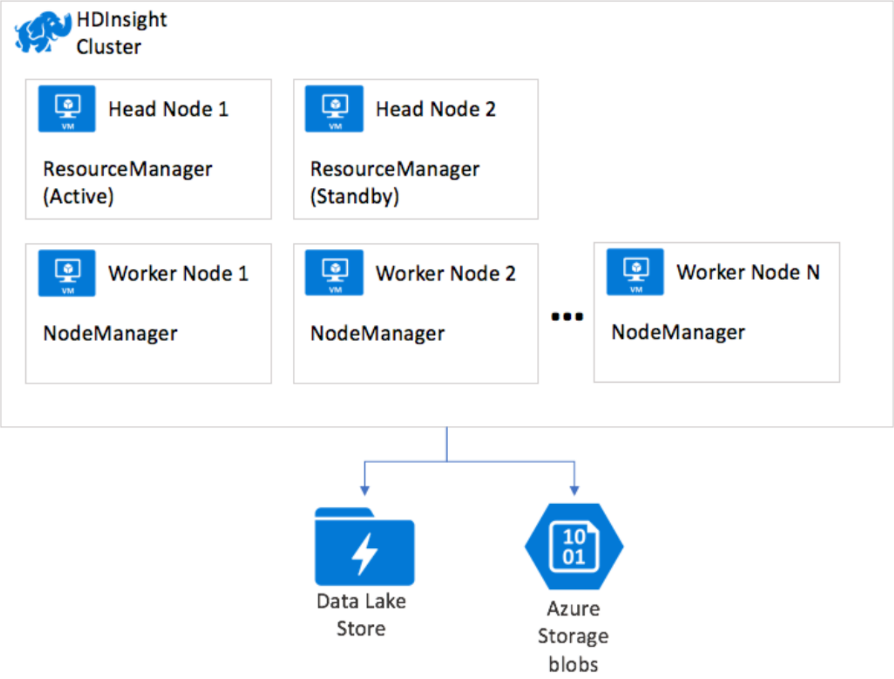
- Named industry leader by Forrester with the most comprehensive, scalable, and integrated platforms*
- Recognized for its cloud-first strategy that is paying off*

*The Forrester WaveTM: Big Data Hadoop Cloud Solutions, Q2 2016.

Hadoop Workloads

YARN on HD Insight

All HDInsight cluster types deploy YARN. The **ResourceManger** is deployed in a high-availability fashion having a primary and secondary instance, which run on the first and second head nodes within the cluster, respectively. Only the one instance of the **ResourceManager** is active at a time. The **NodeManager** instances run across the available Worker Nodes in the cluster.
https://github.com/WilliamAntonRohm/hdinsight-docs/blob/master/hdinsight-architecture.md



HDInsight Offered Cluster Types

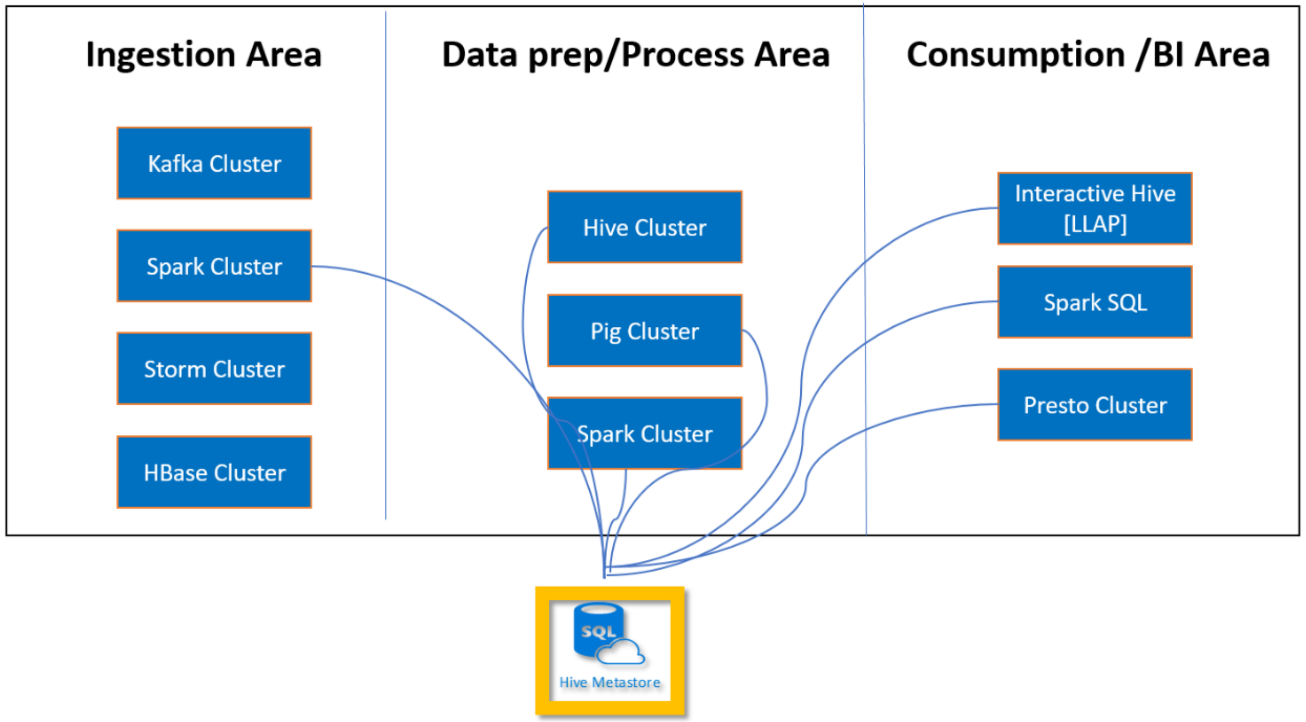
- **Apache Hadoop**: Uses HDFS, YARN resource management, and a simple MapReduce programming model to process and analyze batch data in parallel.
- **Apache Spark**: A parallel processing framework that supports in-memory processing to boost the performance of big-data analysis applications, Spark works for SQL, streaming data, and machine learning. See What is Apache Spark in HDInsight?
- **Apache HBase**: A NoSQL database built on Hadoop that provides random access and strong consistency for large amounts of unstructured and semi-structured data - potentially billions of rows times millions of columns. See What is HBase on HDInsight?
- **Microsoft R Server**: A server for hosting and managing parallel, distributed R processes. It provides data scientists, statisticians, and R programmers with on-demand access to scalable, distributed methods of analytics on HDInsight. See Overview of R Server on HDInsight.
- **Apache Storm**: A distributed, real-time computation system for processing large streams of data fast. Storm is offered as a managed cluster in HDInsight. See Analyze real-time sensor data using Storm and Hadoop.
- **Apache Interactive Hive preview (AKA: Live Long and Process)**: In-memory caching for interactive and faster Hive queries. See Use Interactive Hive in HDInsight.
- **Apache Kafka**: An open-source platform used for building streaming data pipelines and applications. Kafka also provides message-queue functionality that allows you to publish and subscribe to data streams. See Introduction to Apache Kafka on HDInsight.

Components and utilities on HDInsight

- **Ambari**: Cluster provisioning, management, monitoring, and utilities.
- **Avro** (Microsoft .NET Library for Avro): Data serialization for the Microsoft .NET environment.
- **Hive & HCatalog**: SQL-like querying, and a table and storage management layer.
- **Mahout**: For scalable machine learning applications.
- **MapReduce**: Legacy framework for Hadoop distributed processing and resource management.
- **Oozie**: Workflow management.
- **Phoenix**: Relational database layer over HBase.
- **Pig**: Simpler scripting for MapReduce transformations.
- **Sqoop**: Data import and export.
- **Tez**: Allows data-intensive processes to run efficiently at scale.
- **YARN**: Resource management that is part of the Hadoop core library.
- **ZooKeeper**: Coordination of processes in distributed systems.

Using External Metadata Stores

Hive Metastore is critical part of Hadoop architecture as it acts as a central schema repository which can be used by other access tools like Spark, Interactive Hive (LLAP), Presto, Pig, and many other Big Data engines.



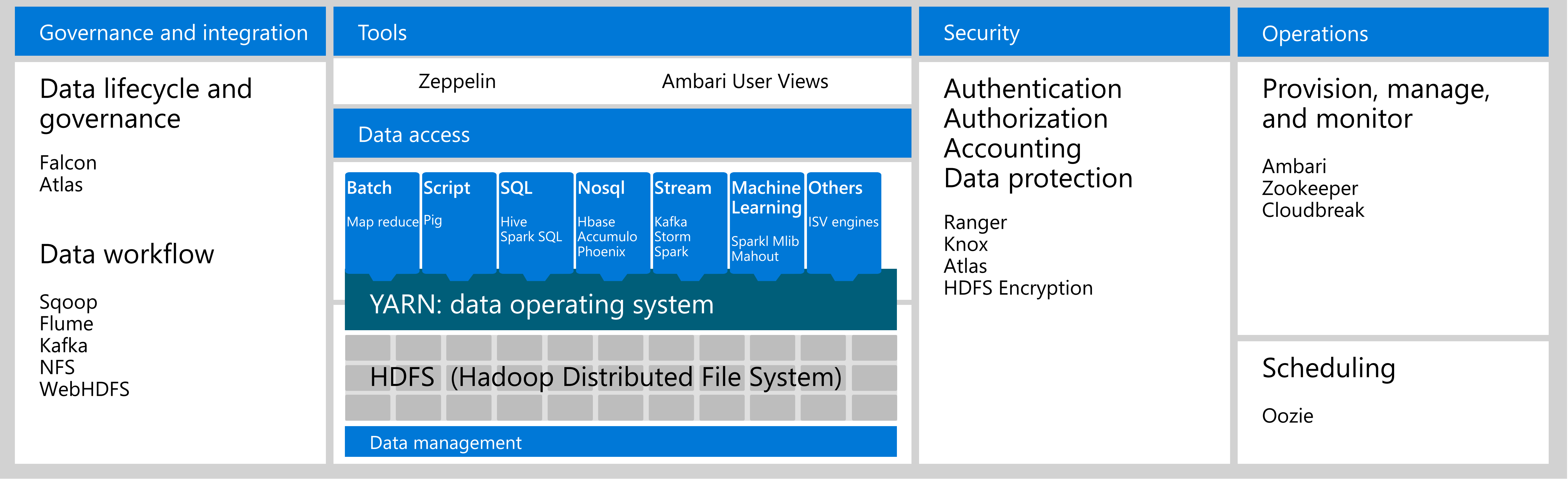
Hadoop is a platform with portfolio of projects

Governed by Apache Software Foundation (ASF)

Comprises core services of MapReduce, HDFS, and YARN

In addition to the core, includes functions across:

Governance and integration, Tools, Data Access, Security, and Operations



Governance and integration

Data lifecycle and governance

Falcon
Atlas

Data workflow

Sqoop
Flume
Kafka
NFS
WebHDFS

Tools

Zeppelin

Ambari User Views

Data access

Batch
Map reduce

Script
Pig

SQL
Hive
Spark SQL

Nosql
Hbase
Accumulo
Phoenix

Stream
Kafka
Storm
Spark

Machine Learning
Sparkl Mlib
Mahout

Others
ISV engines

YARN: data operating system

HDFS (Hadoop Distributed File System)

Data management

Security

Authentication
Authorization
Accounting
Data protection

Ranger
Knox
Atlas
HDFS Encryption

Operations

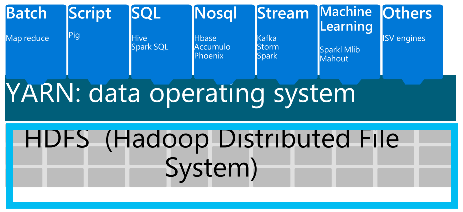
Provision, manage, and monitor

Ambari
Zookeeper
Cloudbreak

Scheduling

Oozie

HDFS



HDFS is a distributed file system

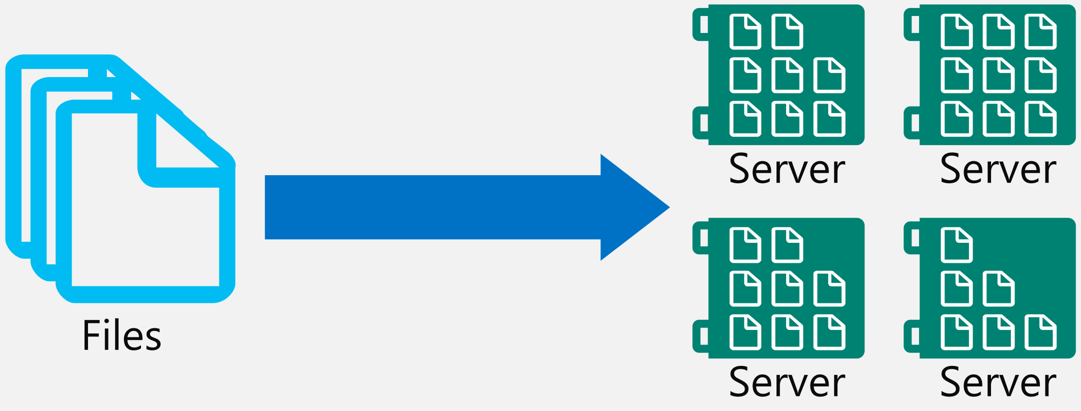
From a few nodes to thousands of nodes

Files can be spread out over multiple nodes

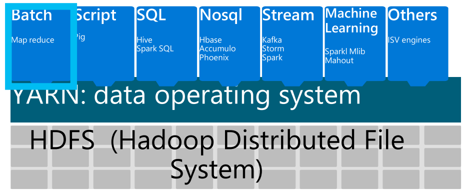
HDFS stores large amounts of data

Very large files are supported including those larger than the capacity of a single node

HDFS stores non-relational files



MapReduce



Takes processing to where data is

Distributed processing: instead of serializing processing through one pipe, distributes computing locally where data is

Brings back only the resultant data

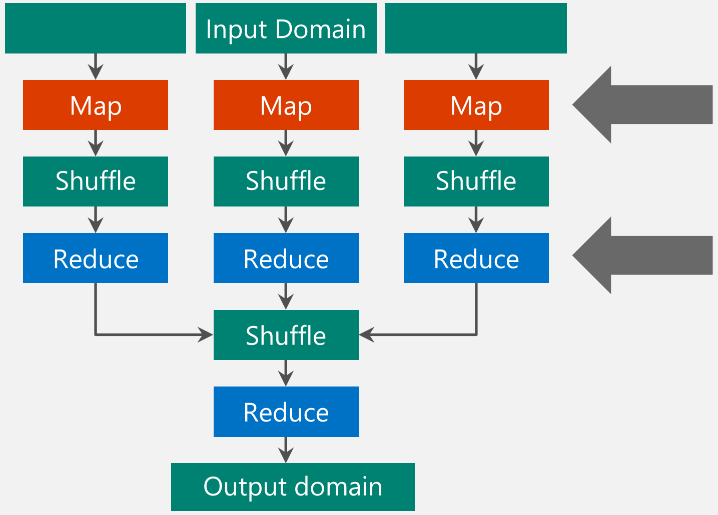
Scales linearly as you add nodes

Three-step execution

Map: Developer writes map functions to the data

Shuffle / Distributes: Framework automatically shuffles for you (networking, synchronization, recovery, scheduling)

Reduce: Developer writes reduce functions to bring resultant data back

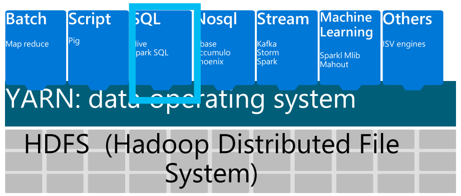


```
// Map Reduce function in JavaScript

var map = function (key, value, context) {
var words = value.split(/[^a-zA-Z]/);
for (var i = 0; i < words.length; i++) {
              if (words[i] !== "")
context.write(words[i].toLowerCase(),
1);}
}};

var reduce = function (key, values, context) {
var sum = 0;
while (values.hasNext()) {
sum += parseInt(values.next());
       }
context.write(key, sum);
};
```

Hive



SQL-like queries on Hadoop data in HDFS

HiveQL is a SQL-like language (subset of SQL)

Hive structures include well-understood database concepts such as tables, rows, columns, partitions

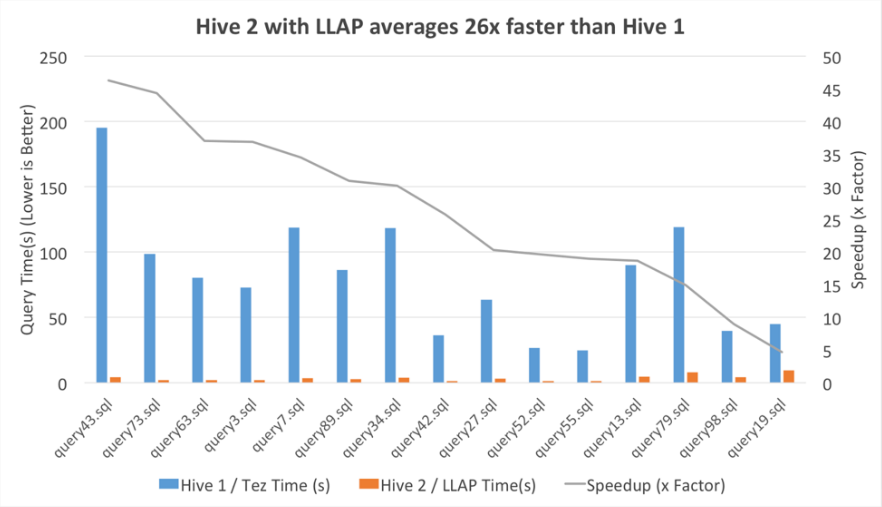
Compiled into MapReduce jobs that are executed on Hadoop

Dramatic performance gains with Hive w/LLAP

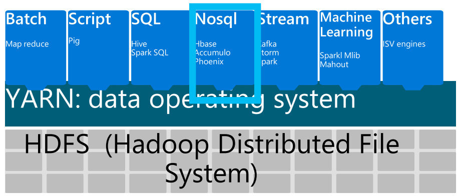
Performance gains up to 25x

ODBC drivers to integrate with Power BI, Tableau, Qlik, etc.

Opens up scenarios to do interactive BI and reporting on big data



HBase

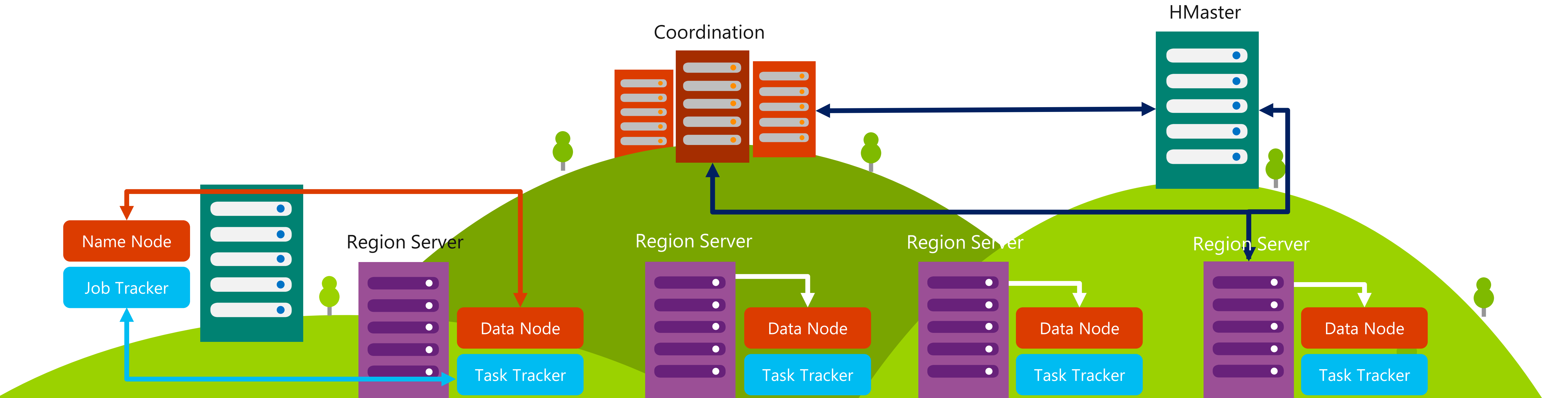


NoSQL database on data in HDFS

Columnar, NoSQL database

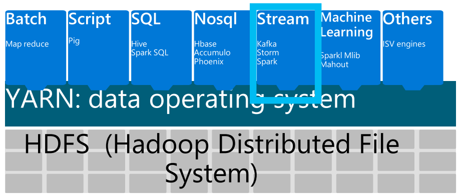
Runs on top of the Hadoop Distributed File System (HDFS)

Provides flexibility in that new columns can be added to column families at any time



Storm

STORM



Batch
Map reduce

Script
Pig

SQL
Hive
Spark SQL

Nosql
Hbase
Accumulo
Phoenix

Stream
Kafka
Storm
Spark

Machine Learning
Sparkl Mlib
Mahout

Others
ISV engines

YARN: data operating system

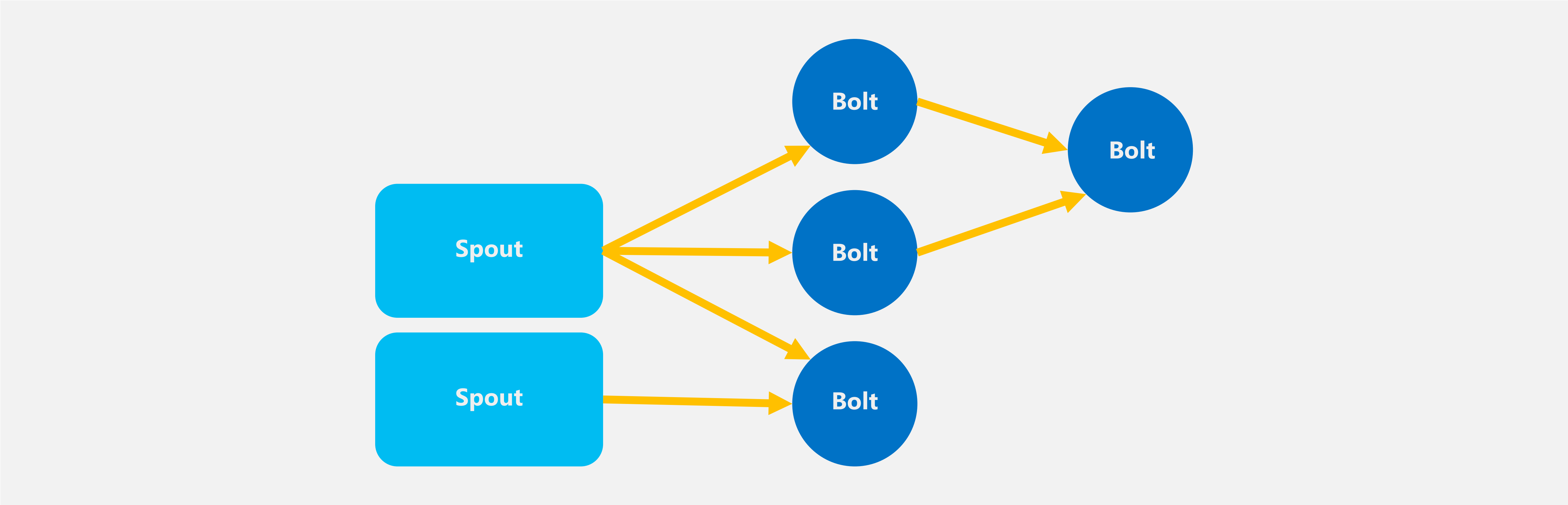
HDFS (Hadoop Distributed File System)

Stream analytics for Near-Real Time processing

Consumes millions of real-time events from a scalable event broker (i.e.; Apache Kafka, Azure Event Hub)

Performs time-sensitive computation

Output to persistent stores, dashboards or devices



Spout

Spout

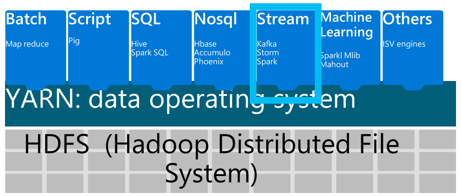
Bolt

Bolt

Bolt

Bolt

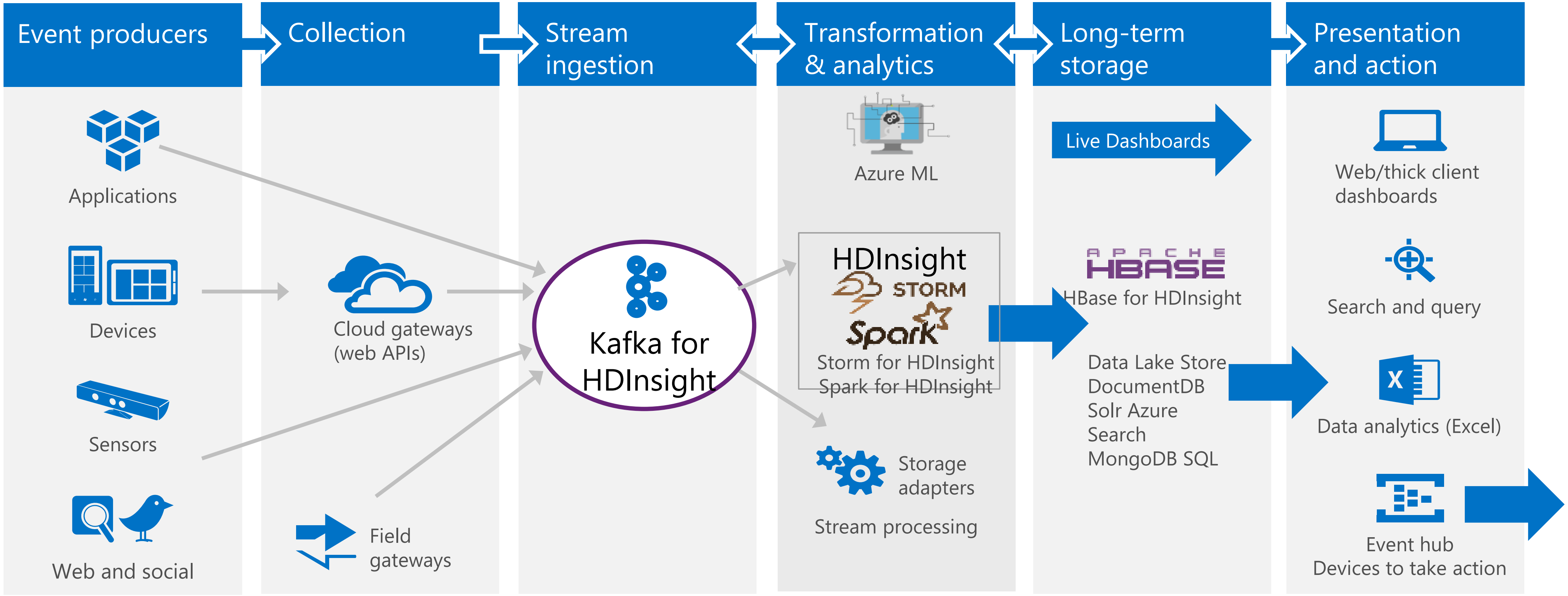
Kafka



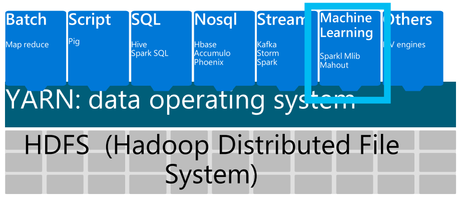
High-throughput, low-latency for real-time data

Stream millions of events per second

Enterprise-grade management and control



Mahout

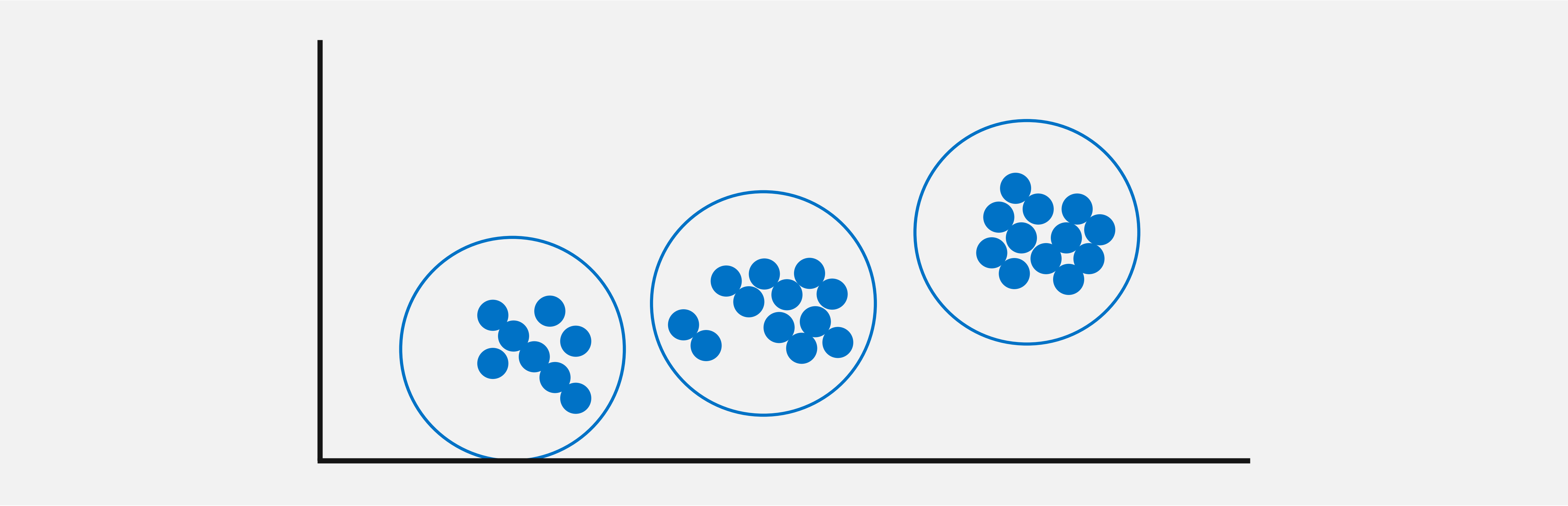


Machine learning library

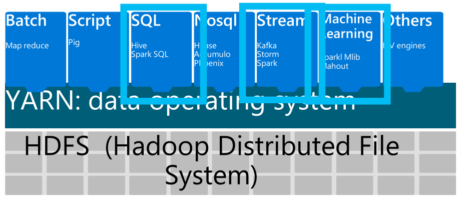
A library of machine learning algorithms to execute on data in HDFS

Algorithms are not dependent on size of data and can scale with large datasets

Library includes: Collaborative Filtering, Classification, Clustering, Dimensionality Reduction, Topic Models



Spark



Massive data processing framework built on in-memory

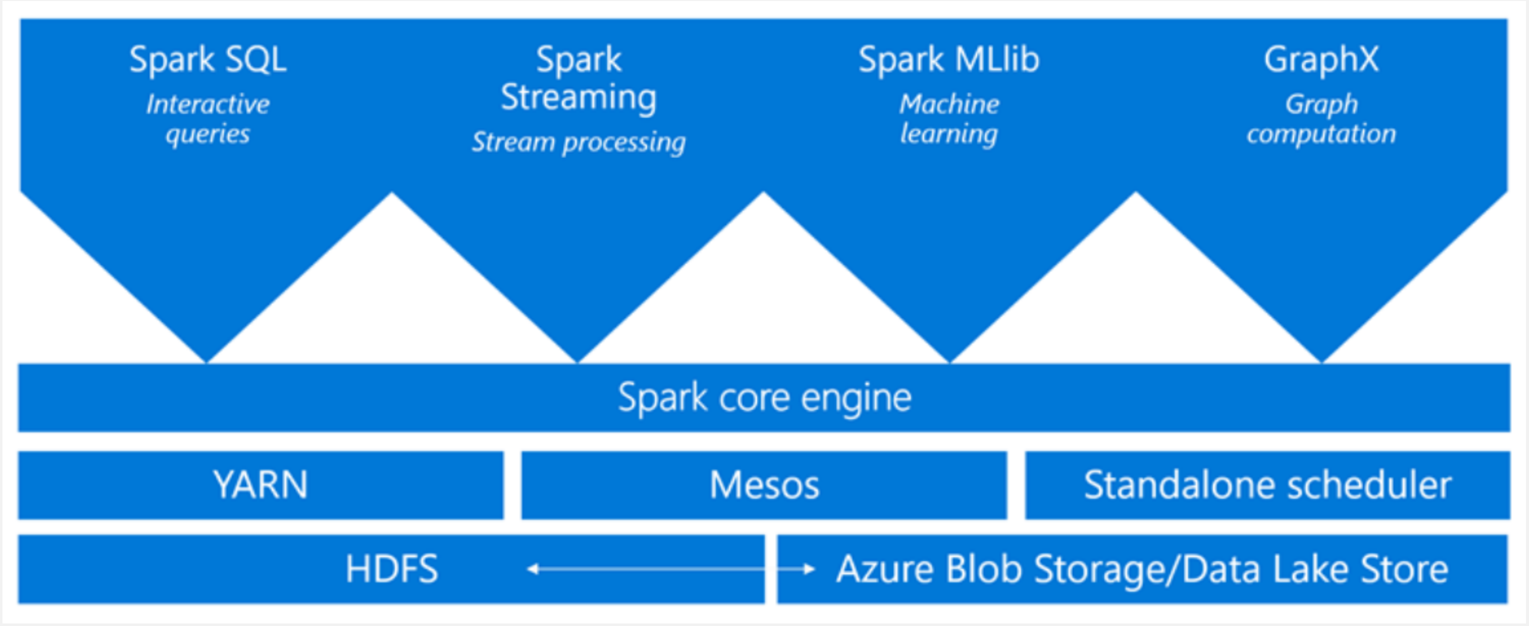
Single execution model for multiple tasks

Processing up to 100x faster performance

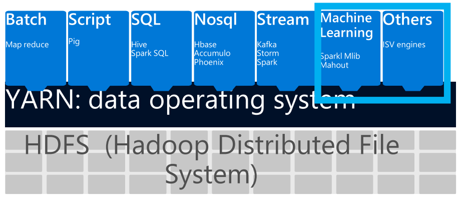
Developer friendly (Java, Python, Scala)

BI tool of choice (Power BI, Tabelau, Qlik, SAP)

Notebook experience (Jupyter & Zeppelin)



R Server



Batch

Map reduce

Script

Pig

SQL

Hive
Spark SQL

Nosql

Hbase
Accumulo
Phoenix

Stream

Kafka
Storm
Spark

Machine Learning

Sparkl Mlib
Mahout

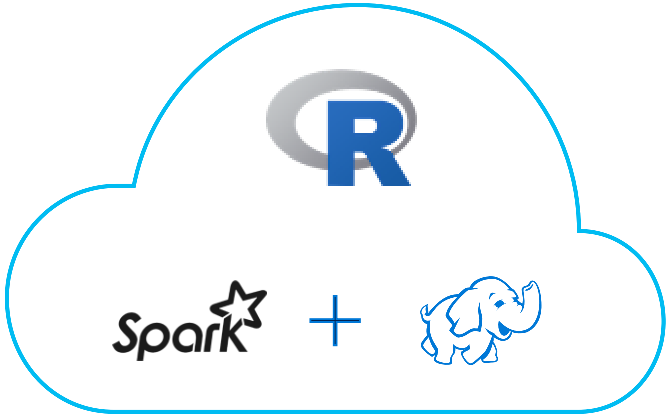
Others

ISV engines

YARN: data operating system

HDFS (Hadoop Distributed File System)

Predictive analytics, machine learning, and statistical modeling for big data

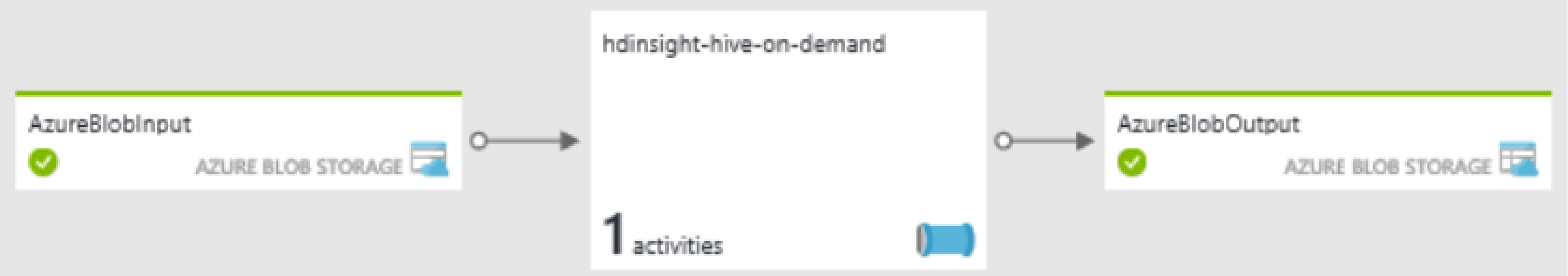


- Largest portable R parallel analytics library
- Terabyte-scale machine learning—1,000x larger than in open source R
- Up to 100x faster performance using Spark and optimized vector/math libraries
- Enterprise-grade security and support

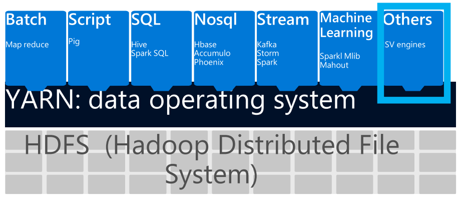
on-demand Hadoop clusters in HDInsight

- You only pay for the time a job is running on the HDInsight Hadoop cluster (plus a brief configurable idle time). The billing for HDInsight clusters is pro-rated per minute, whether you are using them or not. When you use an on-demand HDInsight linked service in Data Factory, the clusters are created on-demand, and the clusters are deleted automatically when the jobs are completed. Therefore, you only pay for the job running time and the brief idle time (time-to-live setting).
- You can create a workflow using a Data Factory pipeline. For example, you can have the pipeline to copy data from an on-premises SQL Server to an Azure blob storage, process the data by running a Hive script and a Pig script on an on-demand HDInsight Hadoop cluster. Then, copy the result data to an Azure SQL Data Warehouse for BI applications to consume.
- You can schedule the workflow to run periodically (hourly, daily, weekly, monthly, etc.).

https://github.com/WilliamAntonRohm/hdinsight-docs/blob/master/hdinsight-hadoop-create-linux-clusters-adf.md



ISV Integration



Batch
Map reduce
Script
Pig
SQL
Hive
Spark SQL
Nosql
Hbase
Accumulo
Phoenix
Stream
Kafka
Storm
Spark
Machine Learning
Sparkl Mlib
Mahout
Others
SV engines
YARN: data operating system
HDFS (Hadoop Distributed File System)

Integration with leading productivity applications

Spin up Hadoop and Spark clusters pre-integrated and pre-tuned with ISV applications out-of-the-box

Runs on the HDInsight clusters; does not require separate VMs

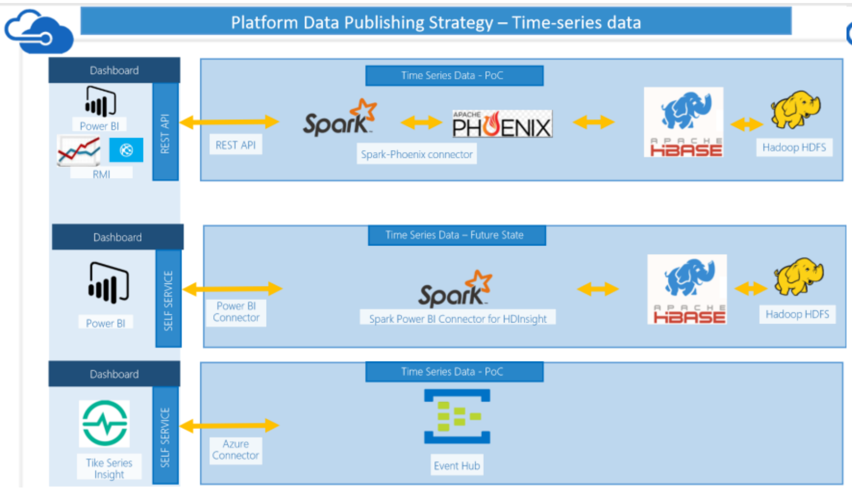
Fast and easy way to spin up applications

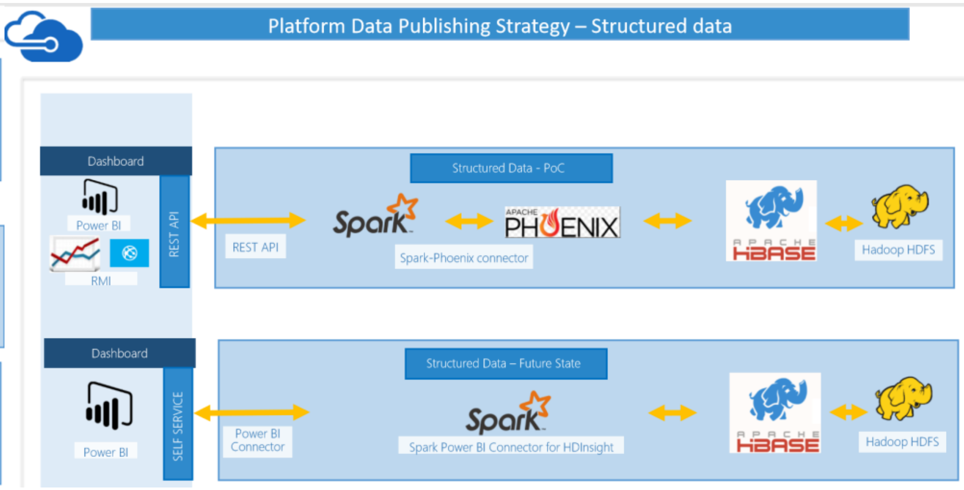
atscale

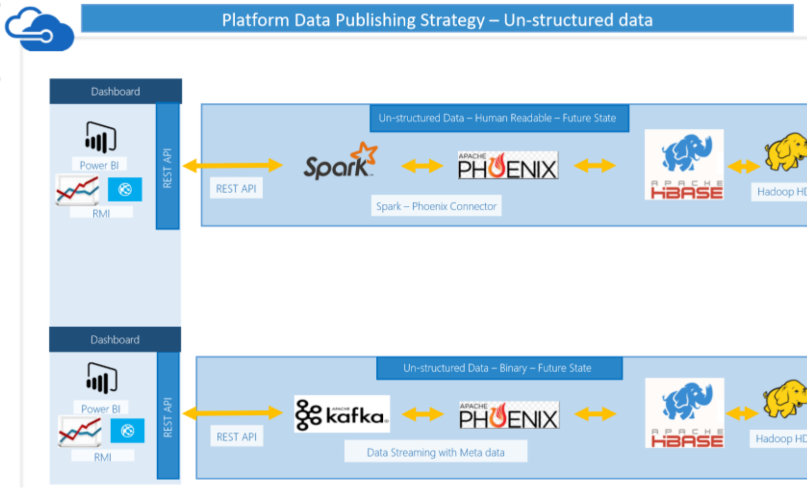
StreamSets™

Datameer

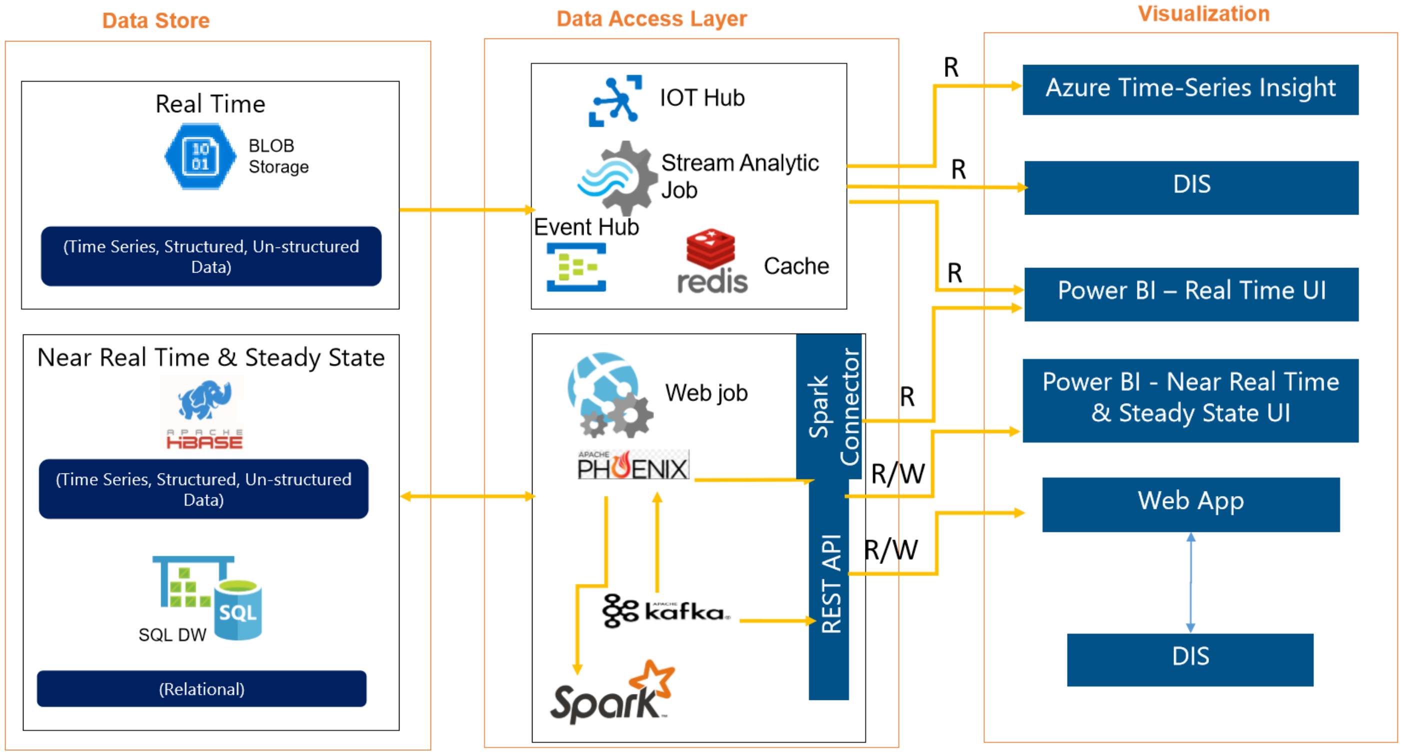
CASK






Platform Data Publishing Strategy – Un-structured data
Dashboard
Power BI
RMI
REST API
Un-structured Data – Human Readable – Future State
REST API
Spark
PHOENIX
Spark – Phoenix Connector
HBASE
Hadoop HD
Dashboard
Power BI
RMI
REST API
Un-structured Data – Binary – Future State
REST API
kafka
PHOENIX
Data Streaming with Meta data
HBASE
Hadoop HD

Data Access Across Real Time, Near Real Time & Steady State Tiers



Virginia Tech is able to capture data from DNA sequencers which are generating **15 PB of genome data each year**. Rather than creating a supercomputing center with millions of dollars, Virginia Tech leverages Azure and only paying for compute they use.

“What excites me about what I’m doing with HDInsight is the ability to accelerate discovery to the point that we may be able to find treatments for cancer.”

Wu Feng
Professor of Computer Science
Virginia Tech

VirginiaTech.

Blackball uses HDInsight to collect point-of-sale (POS) data and new types of data such as customer feedback via social media.

"Before, we thought that people would choose cold drinks and desserts in hot weather. But contrary to our assumptions, in certain outlets we saw an opposite trend."

Andrew Cheong
Senior Manager
BlackBall

黑丸 BlackBall™
Original Taiwanese Tea & Dessert

Get Started

Read documentation

http://azure.microsoft.com/en-us/documentation/services/hdinsight/

Learning Map

http://azure.microsoft.com/en-us/documentation/articles/hdinsight-learn-map/

Microsoft Virtual Academy

http://www.microsoftvirtualacademy.com/training-courses/getting-started-with-microsoft-big-data

Channel 9 Data Exposed Show

http://channel9.msdn.com/Shows/Data-Exposed

Try 30 day trial

http://azure.microsoft.com/en-us/pricing/free-trial/