Microsoft Azure - Starter Kits for Partners

Architecture

Intelligent Apps & Analytics

Last Update: August 2016





**MICROSOFT MAKES NO WARRANTIES, EXPRESS, IMPLIED OR STATUTORY, AS TO THE INFORMATION IN THIS DOCUMENT.**

The information contained in this document represents the current view of Microsoft Corporation on the issues discussed as of the date of publication. Because Microsoft must respond to changing market conditions, it should not be interpreted to be a commitment on the part of Microsoft, and Microsoft cannot guarantee the accuracy of any information presented after the date of publication.

Complying with all applicable copyright laws is the responsibility of the user. Without limiting the rights under copyright, no part of this document may be reproduced, stored in or introduced into a retrieval system, or transmitted in any form or by any means (electronic, mechanical, photocopying, recording, or otherwise), or for any purpose, without the express written permission of Microsoft Corporation.

Microsoft may have patents, patent applications, trademarks, copyrights, or other intellectual property rights covering subject matter in this document. Except as expressly provided in any written license agreement from Microsoft, the furnishing of this document does not give you any license to these patents, trademarks, copyrights, or other intellectual property.

The descriptions of other companies’ products in this document, if any, are provided only as a convenience to you. Any such references should not be considered an endorsement or support by Microsoft. Microsoft cannot guarantee their accuracy, and the products may change over time. Also, the descriptions are intended as brief highlights to aid understanding, rather than as thorough coverage. For authoritative descriptions of these products, please consult their respective manufacturers.

© 2015 Microsoft Corporation. All rights reserved. Any use or distribution of these materials without express authorization of Microsoft Corp. is strictly prohibited.

Microsoft and Windows are either registered trademarks of Microsoft Corporation in the United States and/or other countries.

The names of actual companies and products mentioned herein may be the trademarks of their respective owners.

Contents

[Overview 5](#_Toc459729261)

[This Starter Kits Selected Architecture 5](#_Toc459729262)

[Intelligent Apps 5](#_Toc459729263)

[Information management 6](#_Toc459729264)

[Big data stores 6](#_Toc459729265)

[Machine learning and advanced analytics 6](#_Toc459729266)

[Dashboards and visualizations 6](#_Toc459729267)

[Intelligence 6](#_Toc459729268)

[Solution Architecture 6](#_Toc459729269)

[Data integration and Information Management 8](#_Toc459729270)

[Azure IoT Hub 8](#_Toc459729271)

[Azure Event Hub 9](#_Toc459729272)

[Azure Data Factory 10](#_Toc459729273)

[Big Data Stores 10](#_Toc459729274)

[Azure Data Lake Store 10](#_Toc459729275)

[Azure SQL Data Warehouse 11](#_Toc459729276)

[Machine learning and Advanced analytics 12](#_Toc459729277)

[Azure Machine Learning 13](#_Toc459729278)

[Azure HDInsight 14](#_Toc459729279)

[HDFS storage in HDInsight 15](#_Toc459729280)

[Azure Stream Analytics 17](#_Toc459729281)

[Data visualization with PowerBI 18](#_Toc459729282)

[Intelligence with Cognitive Services 18](#_Toc459729283)

Architecture for Intelligent Apps & Analytics on Azure

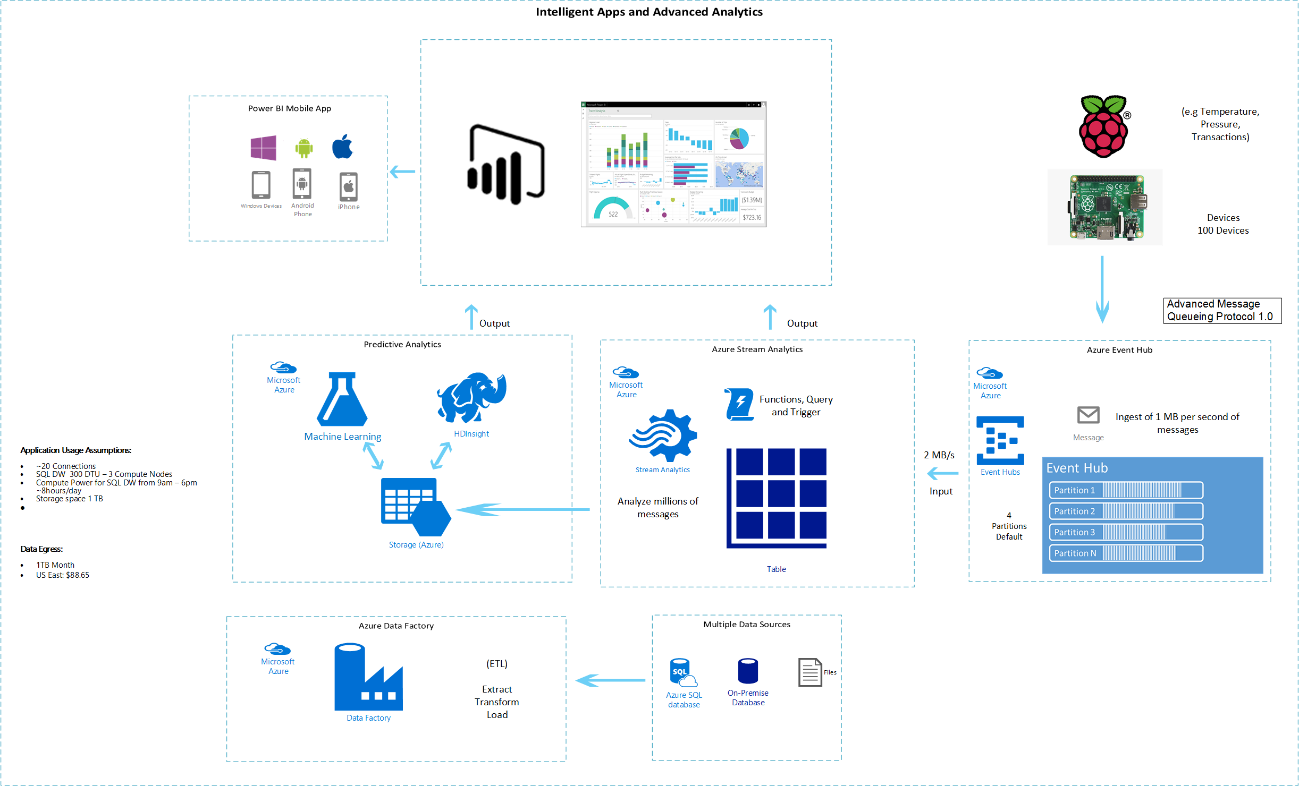
# Overview

The purpose of this document is to complement the lab document to include a broader discussion of concepts relevant to setting up an intelligent app & analytics solution in Azure.

# This Starter Kits Selected Architecture

The scenario below is illustrated in the Cost Calculator and implemented in the Deployment Guidance (Hand on Labs). However, you may make changes in the architecture and topology as appropriate.

This scenario is purely based on Platform as a Service features. It provides higher agility, scalability and elasticity, which is key for migrating existing new applications to Microsoft Azure.



The rest of this document will discuss the other possible architecture aspects for this scenario.

# Intelligent Apps

Microsoft Azure’s data platform and predictive analytics capabilities empower organizations to maximize the value of the data to build intelligent apps. An intelligent app enhance business with machine intelligence to evolve from simple descriptive analytics to prescriptive analytical recommendation, it allow us to predict what will happen with state-of-the-art machine learning, to augment and improve the decision making processes with proactive alerting and prescriptive analytical recommendations on actions to take, to simplify and automate decision making when dealing with complex problems that involve multiple variables that may change in real time. Microsoft Azure brings a set of services to build an integrated intelligent application, these services can be classified into following categories:

# Information management

Orchestrate data movement on a fully managed, end-to-end platform. Use Azure Data Factory to build pipelines and collect and orchestrate data from the services for easier analysis. Plus, use Azure Data Catalog to effectively manage data sources and Azure Event Hubs to provide a staging area for incoming streaming data.

# Big data stores

Store and manage structured data using Azure SQL Data Warehouse that elastically scales with massively parallel processing. Implement a hyper-scale repository with no file size limits for unstructured data using Azure Data Lake Store to attain massive throughput and analytic performance.

# Machine learning and advanced analytics

Design and publish predictive models with Azure Machine Learning, use Azure HDInsight to analyze data in Storm and Spark for Hadoop environments, integrate the code from R or Python, and analyze any kind or any size of data with Azure Data Lake Analytics and Azure Stream Analytics.

# Dashboards and visualizations

Transform data into rich visuals to organize and share so that we can focus on what matters to us. PowerBI is a suite of business analytics tools to analyze data and share insights with rich dashboards available on every device.

# Intelligence

Cognitive Services enable natural and contextual interactions in the apps. Integrate analytics services and models with Cortana to let users interact with the app through speech and receive proactive notifications. Build and connect intelligent bots that naturally interact with apps’ users wherever they are—from SMS to Skype and Office 365—by using the Bot Framework.

# Solution Architecture

The solution architecture descripted in this document is purely base on the Azure PaaS services. PaaS is a complete development and deployment environment in the cloud, with resources that build everything from simple cloud-based apps to sophisticated, cloud-enabled enterprise applications. Developer purchase the resources from a cloud service provider on a pay-as-you-go basis and access them over a secure Internet connection.

Like IaaS, PaaS includes infrastructure—servers, storage, and networking—but also middleware, development tools, business intelligence (BI) services, database management systems, and more. PaaS is designed to support the complete web application lifecycle: building, testing, deploying, managing, and updating.

PaaS allows developers to avoid the expense and complexity of buying and managing software licenses, the underlying application infrastructure and middleware or the development tools and other resources. Developers manage the applications and services, while the cloud service provider typically manages everything else.

With the capabilities that Azure PaaS services provides, we can solve a specific and commonly occurring architecture problems that may be encountered when build this intelligent application, such as Performance and Scalability, Availability, Security, Data management and Messaging.

Performance and Scalability

Performance is an indication of the responsiveness of a system, while scalability is ability of a system either to handle increases in load without impact on performance or for the available resources to be readily increased. Application should be able to scale out within limits to meet peaks in demand, and scale in when demand decreases.

Availability

Availability defines the proportion of time that the system is functional and working. It is usually measured as a percentage of uptime. Azure services provide users with a service level agreement, and allow the application to be designed and implemented in a way that maximizes availability.

Security

Security is the capability of a system to prevent malicious or accidental actions outside of the designed usage, and to prevent disclosure or loss of information. Azure PaaS services provide many options the enable the application can be designed and deployed in a way that protects them from malicious attacks, restricts access to only approved users, and protects sensitive data.

Data management

Data management is the key element of cloud applications, Data is typically hosted in different locations and across multiple servers for reasons such as performance, scalability or availability, and this can present a range of challenges. Azure data services help us to management the data more easily and more securely.

Messaging

The distributed nature of cloud applications requires a messaging infrastructure that connects the components and services, ideally in a loosely coupled manner in order to maximize scalability. Asynchronous messaging is widely used, and provides many benefits, but also brings challenges such as the ordering of messages, poison message management, idempotency, and more.

RESTful Interfaces

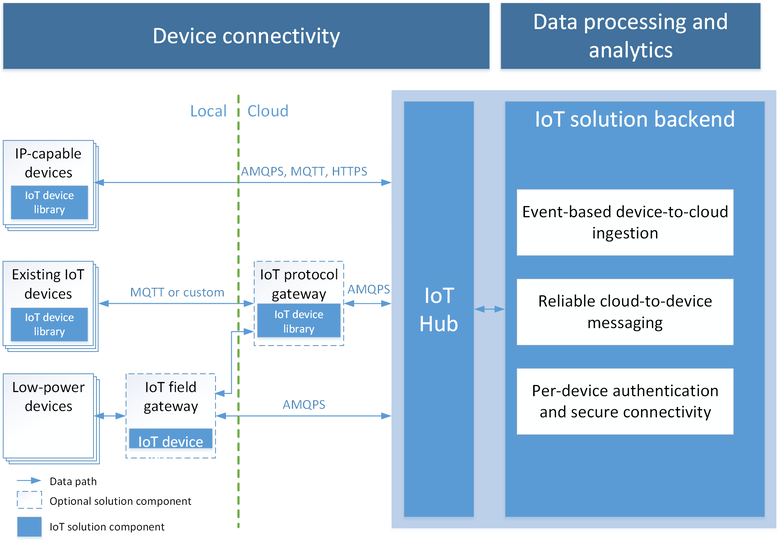
REST is an architectural style of API consisting of a coordinated set of components, connectors and data elements within a distributed system, where the focus is on component roles and a specific set of interactions between data element s rather than implementation details. Microsoft Azure PaaS is characterized by the REST interfaces for a fine grained interaction with the deployment artifacts. All the resources in azure has a REST interfaces that developers can use in the cloud application development to manage the azure resources. Besides, the architecture in this intelligent application is service oriented, the communication between different components are through the well-defined REST interfaces or though the distributed messaging services that azure provides.

# Data integration and Information Management

Data is the fundamental of an intelligent solution, we need to integrate and manage both the relational, non-relational and streaming data from many types of data sources such as web clickstreams, social media, server logs, devices and sensors and geo-location data. Some data may need to be persisted in a repository such as a database or a NoSQL database, while other data may be accessible only as a stream of events. The architecture of our intelligent application has the following data integration and management services on Azure:

## Azure IoT Hub

Azure IoT Hub is a service that provides the capability that an application to collect data from devices and sensors in the Internet of Things(IoT) scenarios. It is a fully managed service that enable reliable and secure bidirectional communications between millions of devices and an application back end.



By using the Azure IoT Hub, the application can implement the data collection with the following communication patterns:

**Event-based device-to-cloud data ingestion:** IoT Hub can receive millions of events per second from the devices and sensors and process them on the hot path by using an event processor engine such as the Azure Stream Analytics or Hadoop Spark. Besides, it can also store them on cold path for analysis. IoT hub retains the event data for up to seven days to guarantee reliable processing and to absorb peaks in the load.

**Reliable cloud-to-device messaging:** The application backend can use IoT Hub to send messages with an at-least-once delivery guarantee to individual devices.

**Upload files and cached sensor data to the cloud:** Devices can upload files to Azure Storage using SAS URIs managed by IoT Hub.

## Azure Event Hub

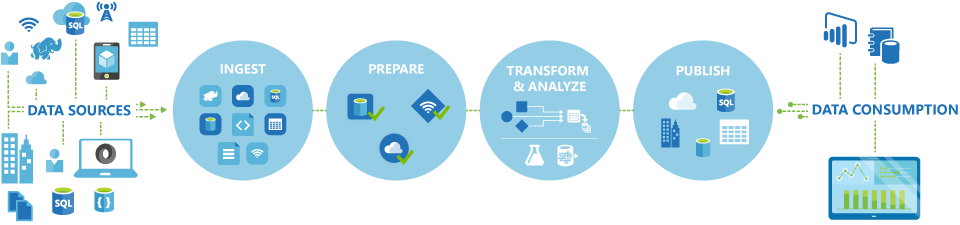
Azure Event Hubs is a highly scalable service that can ingest millions of events per second and stream them into the intelligent application. Application then process and analyze the massive amounts of data produced by the devices or applications. Once the data collected by Event Hubs, then the data will be transformed and stored by using any real-time analytics services or tools.



Azure Event Hubs provides a hyper-scale event and telemetry processing service that can be used for common application and user workflow monitoring at any scale. With the ability to provide publish-subscribe capabilities with low latency and at massive scale, Event Hubs serve as the "on ramp" for Big Data. With publisher-based identity and revocation lists, these capabilities are extended into common Internet of Things scenarios.

## Azure Data Factory

Azure Data Factory is a cloud-based data integration service that orchestrates and automates the movement and transformation of data, it orchestrates existing services that collect raw data and transform it into ready-to-use information.



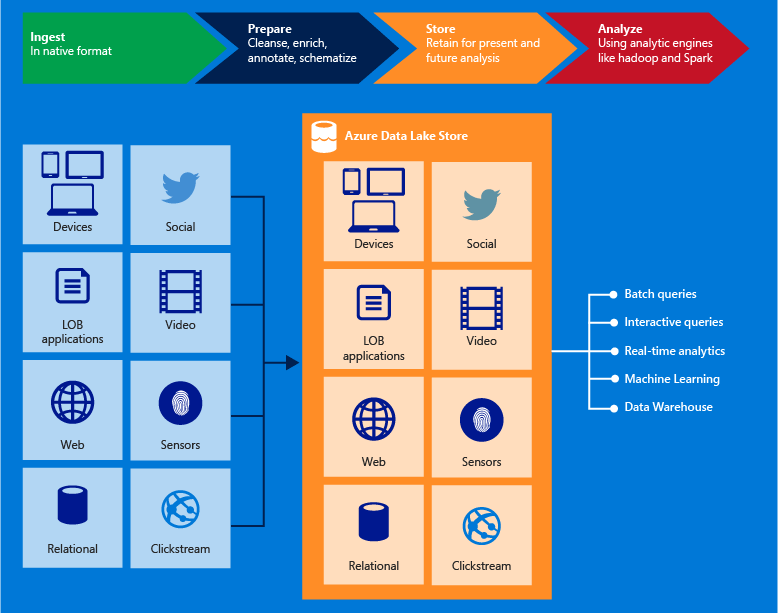
Data Factory works across on-premises and cloud data sources and SaaS to ingest, prepare, transform, analyze, and publish your data. Use Data Factory to compose services into managed data flow pipelines to transform your data using services like Azure HDInsight (Hadoop) and Azure Batch for your big data computing needs, and with Azure Machine Learning to operationalize your analytics solutions. In a word, Data Factory can be used to create highly available data flow pipelines for many scenarios across different scenarios for their analytics pipeline needs.

# Big Data Stores

The application data brought in can be persisted in flexible big data storage service like Azure Data Lake Store and Azure SQL Data Warehouse as a centralized repository for structured and unstructured data with scale for enterprise-wide analytics.

## Azure Data Lake Store

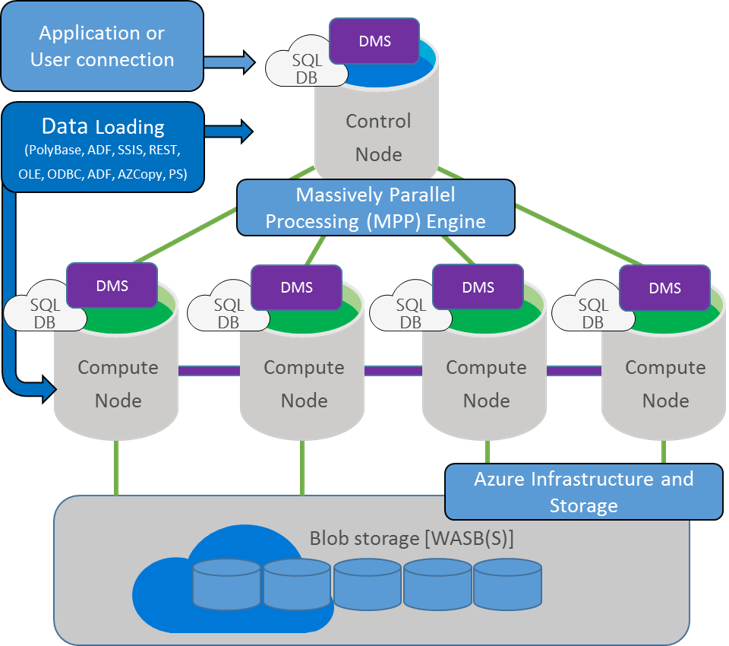
Azure Data Lake Store is an enterprise-wide hyper-scale repository for big data analytic workloads, you can capture data of any size, type, and ingestion speed in one single place for operational and exploratory analytics.



Azure Data Lake Store removes the complexities of ingesting and storing data while making it faster to get up and running with batch, streaming, and interactive analytics. It is specifically designed to enable analytics on the stored data and is tuned for performance for data analytics scenarios, including all the enterprise-grade capabilities—security, manageability, scalability, reliability, and availability—essential for real-world enterprise use scenarios.

## Azure SQL Data Warehouse

Azure SQL Data Warehouse is a cloud-based, scale-out database capable of processing massive volumes of data, both relational and non-relational. Built on our massively parallel processing (MPP) architecture, in this intelligent application, SQL Data Warehouse can store and process the data from different sources.



SQL Data Warehouse is a massively parallel processing (MPP) distributed database system. By dividing data and processing capability across multiple nodes, SQL Data Warehouse can offer huge scalability - far beyond any single system. Behind the scenes, SQL Data Warehouse spreads your data across many shared-nothing storage and processing units. The data is stored in Premium locally redundant storage, and linked to compute nodes for query execution.

SQL Data Warehouse is based on the SQL Server relational database engine, and includes many of the features you expect from an enterprise data warehouse. If you already know T-SQL, it's easy to transfer your knowledge to SQL Data Warehouse.

Using SQL Data Warehouse with PolyBase, we can move data across the ecosystem. PolyBase enable us to leverage the data from different sources by using familiar T-SQL commands. Application can query non-relational data held in Azure Blob storage as though it is a regular table, or import non-relational data into SQL Data Warehouse.

# Machine learning and Advanced analytics

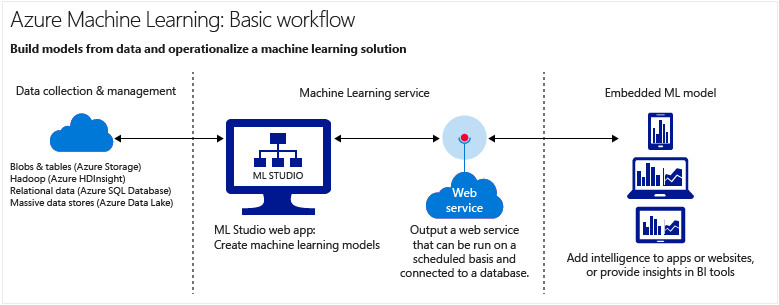
An intelligent app enhance business with machine intelligence to evolve from simple descriptive analytics to prescriptive analytical recommendation. To archive this goal, application need to utilize the predictive analytics to analyze collected or current data for patterns or trends to forecast future events.

Advanced analytics refers to the trends, technologies, and potential for businesses to obtain valuable insight from large, of structured, unstructured, and fast-moving data. The world of data is challenged by increasing volume, variety, velocity and variability of the big data. The volume of data stored is growing exponentially, the variety of data being stored are also becoming more diverse, the velocity of data collected is increasing and need to be able to make real-time analytics based on streaming data, the variability refers to the number of options and variable interpretations which makes data analysis a tough job to do.

Advanced analytics with Azure Big data solution unleashes actionable insights from both structured and unstructured data, it brings the simplicity of Windows to Hadoop through integration with key Microsoft components and offers elastic scale of the cloud through a Hadoop based service on Azure. It is an Enterprise-ready Hadoop based on the Hortonworks Data Platform which offers the most reliable, innovative and trusted distribution available.

## Azure Machine Learning

Azure Machine learning is a powerful way to do predictive analytics. Applications can integrate a set of ready-to-use library of algorithms, create models and deploy the predictive solution quickly. It not only provides tools to model predictive analytics, but also provides a fully-managed service can be used to deploy the predictive models as ready-to-consume web services. The intelligent application then can consume the predictive model web services and provide the data to get the predictive result to make intelligent recommendation.



We can create predictive analytics solutions in the cloud from a large algorithm library, to a studio for building models, to an easy way to deploy the model as a web service. In Machine Learning Studio, predictive models can be created by dragging, dropping and connecting modules.

## Azure HDInsight

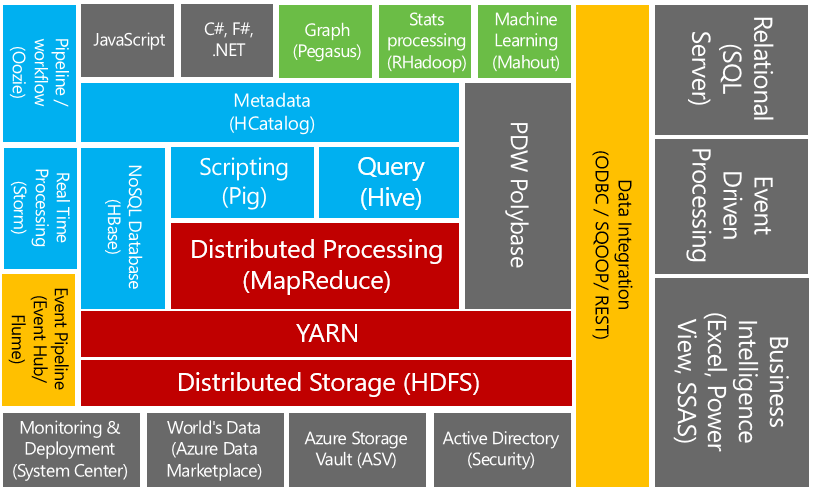
HDInsight is a cloud implementation on Microsoft Azure of the rapidly expanding Apache Hadoop technology stack that is the go-to solution for big data analysis. It includes implementations of Apache Spark, HBase, Storm, Pig, Hive, Sqoop, Oozie, Ambari, and so on. HDInsight also integrates with business intelligence (BI) tools such as Power BI, Excel, SQL Server Analysis Services, and SQL Server Reporting Services.

**Hadoop:** Provides reliable data storage with HDFS, and a simple MapReduce programming model to process and analyze 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.

**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.

**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.



The following components and utilities Ambari: Cluster provisioning, management, monitoring, and utilities.

**Avro:** Data serialization for the Microsoft .NET environment.

**Hive & HCatalog:** Structured Query Language (SQL)-like querying, and a table and storage management layer.

**Mahout:** Machine learning.

**MapReduce:** Legacy framework for Hadoop distributed processing and resource management. See YARN, the next-generation resource framework.

**Oozie:** Workflow management.

**Phoenix:** Relational database layer over HBase.

**Pig:** Simpler scripting for MapReduce transformations.

**Sqoop:** Data import and export.

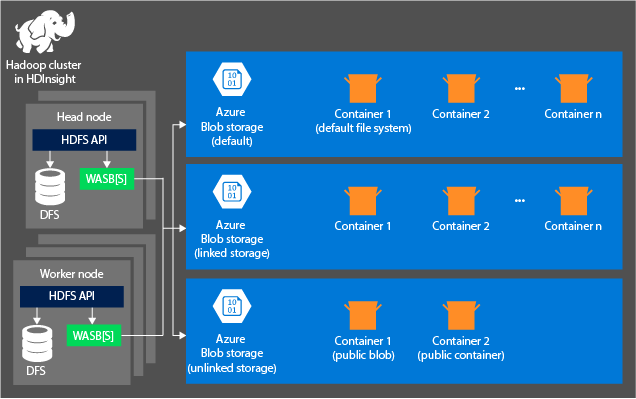
**Tez:** Allows data-intensive processes to run efficiently at scale.

**YARN:** Part of the Hadoop core library and next generation of the MapReduce software framework.

**ZooKeeper:** Coordination of processes in distributed systems. are included on HDInsight clusters:

## HDFS storage in HDInsight

Hadoop Distributed File System(HDFS) is a distributed file system that, with MapReduce and YARN, is the core of the Hadoop components, HDFS is the standard file system for Hadoop clusters on HDInsight. In HDInsight, we can use the low-cost Azure Blob storage as the data storage for HDFS. Through the HDFS interface, the full set of components in HDInsight can operate directly on structured or unstructured data in Blog storage. Following is a abstract view of the HDInsight storage architecture with Azure Blog storage:



Hadoop default file system implies a default scheme and authority. It can be used to resolve relative paths. During the HDInsight creation process, an Azure Storage account and a specific Azure Blob storage container from that account is designated as the default file system.

In addition to this storage account, Additional storage accounts can be added from the same Azure subscription or different Azure subscriptions during the creation process.

There are several benefits associated with storing the data in Azure Blob storage:

**Data reuse and sharing:** The data in Azure Blog storage can be accessed either through the HDFS APIs or through the Blog Storage REST APIs.

**Data archiving:** Storing data in Azure Blob storage enables the HDInsight clusters used for computation to be safely deleted without losing user data.

**Data storage cost:** Storing data in DFS for the long term is more costly than storing the data in Azure Blob storage because the cost of a compute cluster is higher than the cost of an Azure Blob storage container.

**Elastic scale-out:** Although HDFS provides you with a scaled-out file system, the scale is determined by the number of nodes that you create for your cluster. Changing the scale can become a more complicated process than relying on the elastic scaling capabilities of Azure Blob storage.

**Geo-replication:** Azure Blob storage containers can be geo-replicated. Although this gives application geographic recovery and data redundancy.

## Azure Stream Analytics

Azure Stream Analytics is a fully managed, cost effective real-time event processing engine that helps to unlock deep insights from data. Stream Analytics makes it easy to set up real-time analytic computations on data streaming from devices, sensors, web sites, social media, applications, infrastructure systems, and so on.

With the Azure Stream analytics, our intelligent application can process on the real-time streaming data and pass the data into HDInsight for further processing and analytics. Stream analytics provides following key capabilities and benefits:

**Ease of use:** Stream Analytics supports a simple, declarative query model for describing transformations. It uses a T-SQL variant, and removes the need for customers to deal with the technical complexities of stream processing systems.

**Scalability:** Stream Analytics is capable of handling high event throughput of up to 1GB/second. Integration with Azure Event Hubs and Azure IoT Hubs allow the solution to ingest millions of events per second coming from connected devices, clickstreams, and log files, to name a few.

**Reliability, repeatability and quick recovery:** A managed service in the cloud, Stream Analytics helps prevent data loss and provides business continuity in the event of failures through built-in recovery capabilities.

**Low cost:** As a cloud service, Stream Analytics is optimized to provide users a very low cost to get going and maintain real-time analytics solutions. The service is built to pay as you go based on Streaming Unit usage and the amount of data processed by the system.

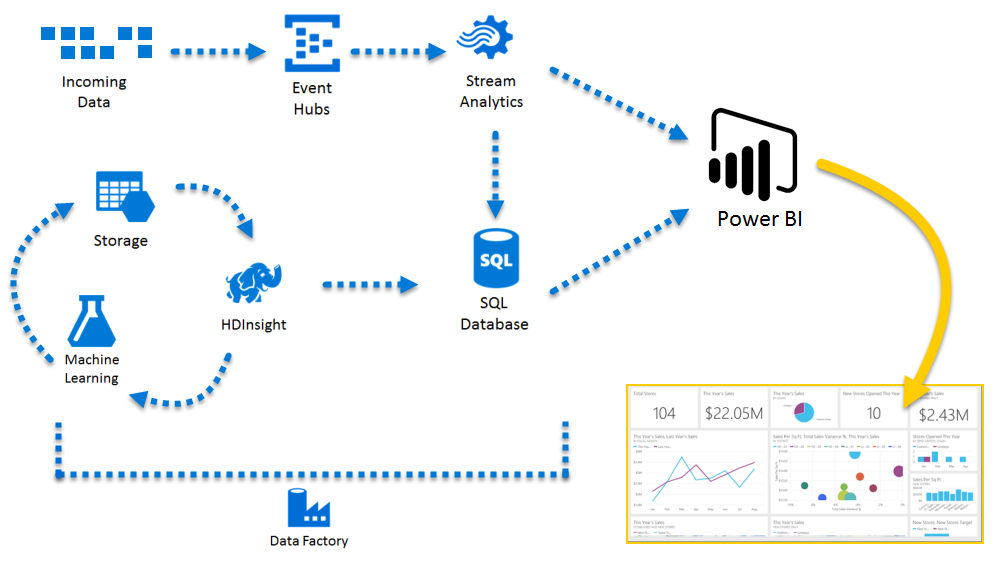
**Reference data:** Stream Analytics provides users the ability to specify and use reference data.

**User Defined Functions:** Stream Analytics has integration with Azure Machine Learning to define function calls in the Machine Learning service as part of a Stream Analytics query. This expands the capabilities of Stream Analytics to leverage existing Azure Machine Learning solutions.

**Connectivity:** Stream Analytics connects directly to Azure Event Hubs and Azure IoT Hubs for stream ingestion, and the Azure Blob service to ingest historical data. Results can be written from Stream Analytics to Azure Storage Blobs or Tables, Azure SQL DB, Azure Data Lake Stores, DocumentDB, Event Hubs, Azure Service Bus Topics or Queues, and Power BI, where it can then be visualized, further processed by workflows, used in batch analytics via Azure HDInsight or processed again as a series of events. When using Event Hubs it is possible to compose multiple Stream Analytics together with other data sources and processing engines without losing the streaming nature of the computations.

# Data visualization with PowerBI

In this intelligent application architecture, we apply the PowerBI to visualize our data and the analysis results. With Azure services and PowerBI, all data processing efforts can be turn into analytics and reports that provide real-time insights into the business. PowerBI has a multitude of Azure connections available, and the business intelligence solutions you can create with those services are as unique as your business. In PowerBI, you can integrate multi-source data processing, make use of massive real-time systems, use Stream Analytics and Event Hubs, and so on.



There are all sorts of scenarios where Azure and Power BI can be combined - the possibilities and opportunities are as unique as your business.

# Intelligence with Cognitive Services

Cognitive Services are a new collection of intelligence and knowledge APIs that enable developers to ultimately build the intelligent apps, and this APIs democratize the artificial intelligence to allow developers with different skill levels be able to use these APIs quickly.

Regardless of the platform, apps can call the Cognitive Services APIs as long as the app has an Internet connection, every apps can integrate it to build a smarter, more context-aware applications. Cognitive Services includes the following categories of APIs:

**Vision:** From faces to feelings, allow your apps to understand images and video.

**Speech:** Hear and speak to your users by filtering noise, identifying speakers, and understanding intent.

**Language:** Process text and learn how to recognize what users want.

**Knowledge:** Tap into rich knowledge amassed from the web, academia, or your own data.

**Search:** Access billions of web pages, images, videos, and news with the power of Bing APIs.

The Cognitive Services is easy to implement because of they are all of simple REST calls, is flexible enough to allow developer be able to find what intelligence features they need and then integrate the intelligence features to the cross-platform apps.