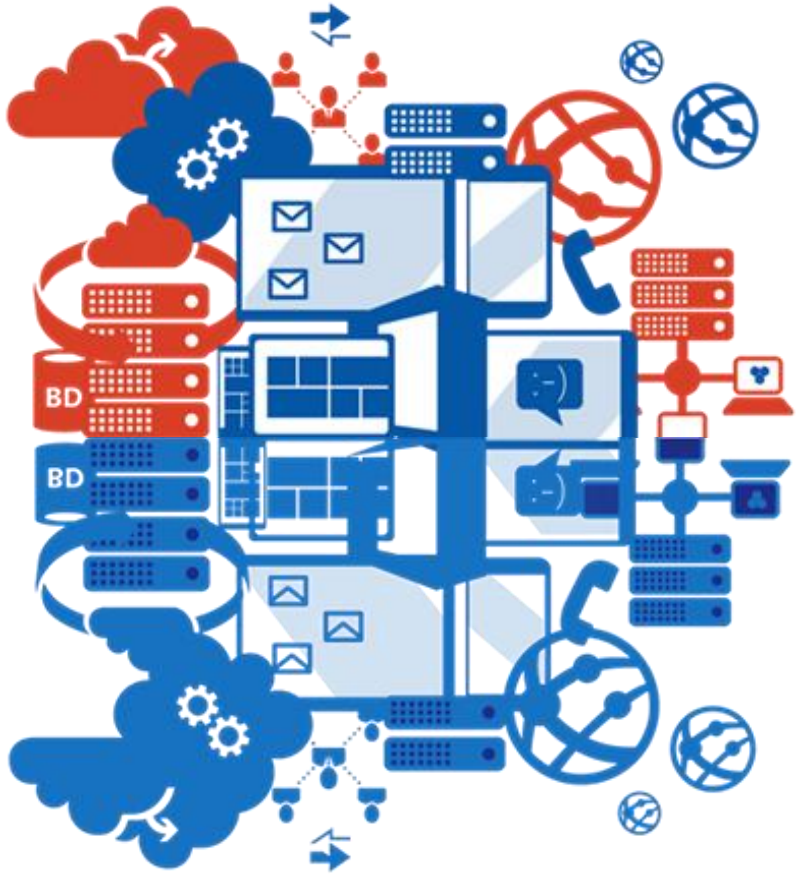
Microsoft

Sparking your Knowledge with Azure Spark



Data Platform Airlift

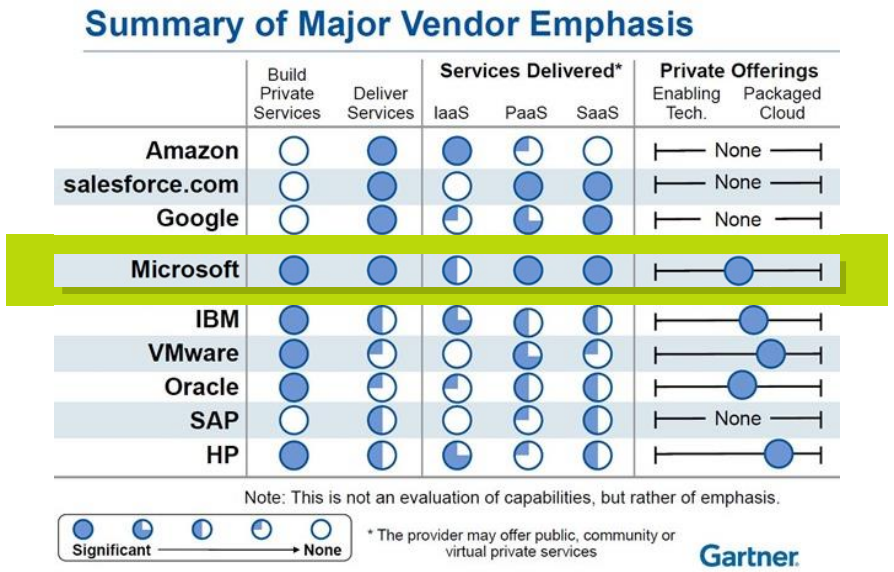
21 de Outubro \\ Microsoft Lisbon Experience

Microsoft Azure

Microsoft

Windows

Industry validation

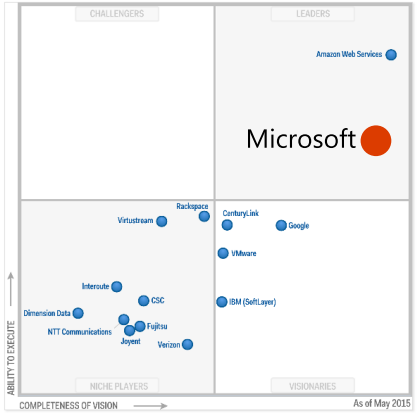


"Microsoft's comprehensive hybrid story, which spans applications and platforms as well as infrastructure, is highly attractive to many companies, drawing them towards the cloud in general."

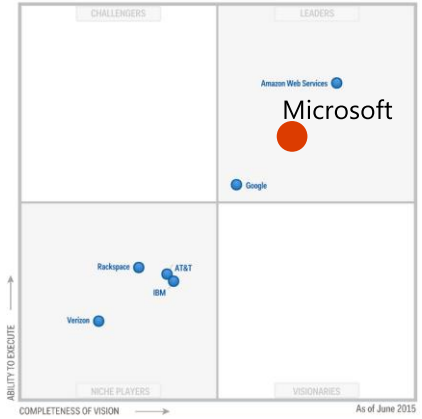
LYDIA LEONG,
GARTNER

Microsoft Leads Everywhere…

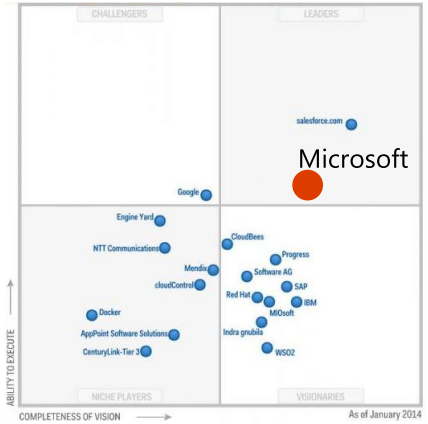
Public Cloud IaaS (May 2015)



Cloud Storage (June 2015)

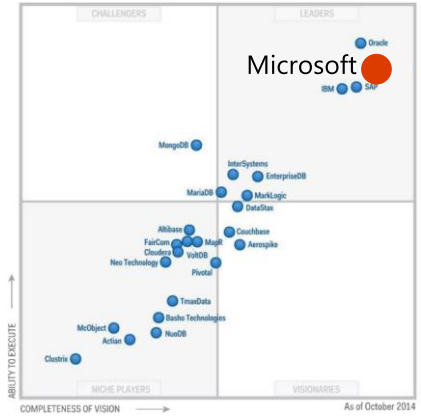


Enterprise App PaaS (Jan 2014)



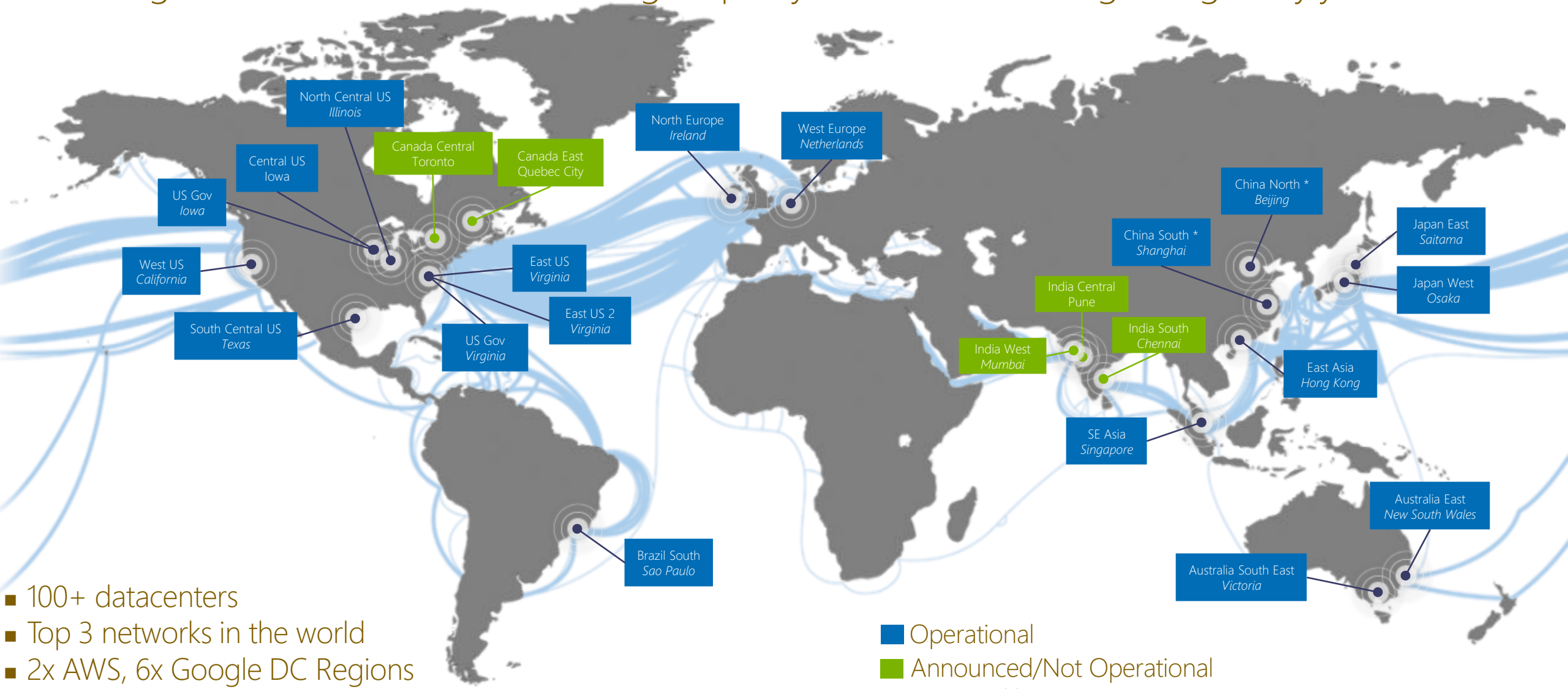
X86 Server Virt (July 2015)

Operational DBMS Systems (Oct 2014)



Huge infrastructure scale is the enabler

24 Regions Worldwide, 19 ONLINE…huge capacity around the world…growing every year



- 100+ datacenters
- Top 3 networks in the world
- 2x AWS, 6x Google DC Regions
- G Series – Largest VM in World, 32 cores, 448GB Ram, SSD…

Operational

Announced/Not Operational

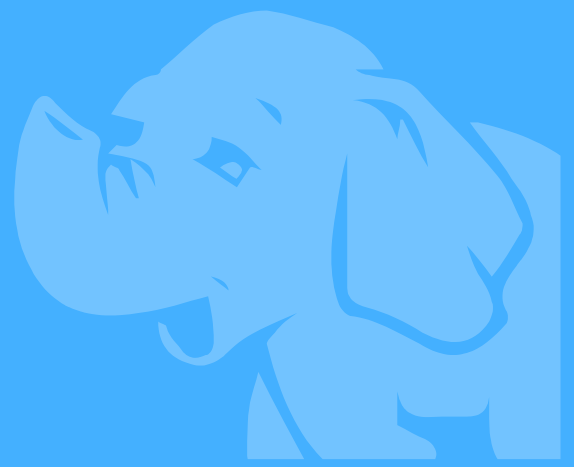
* Operated by 21Vianet

Platform Services

Security & Management

- Portal
- Active Directory
- Multi-Factor Authentication
- Automation
- Key Vault
- Store / Marketplace
- VM Image Gallery & VM Depot

Spark with Azure HDInsight



Hybrid Operations

- Azure AD Connect Health
- AD Privileged Identity Management
- Backup
- Operational Insights
- Import/Export
- Site Recovery
- StorSimple

Infrastructure Services

Compute

- Virtual Machines
- Containers

Storage

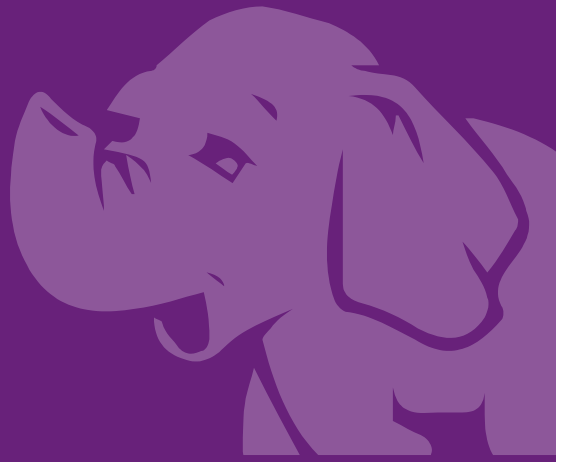
- BLOB Storage
- Azure Files
- Premium Storage

Networking

- Virtual Network
- Load Balancer
- DNS
- Express Route
- Traffic Manager
- VPN Gateway
- Applicat. Gateway

Datacenter Infrastructure (24 Regions, 19 Online)

Apache Spark Overview

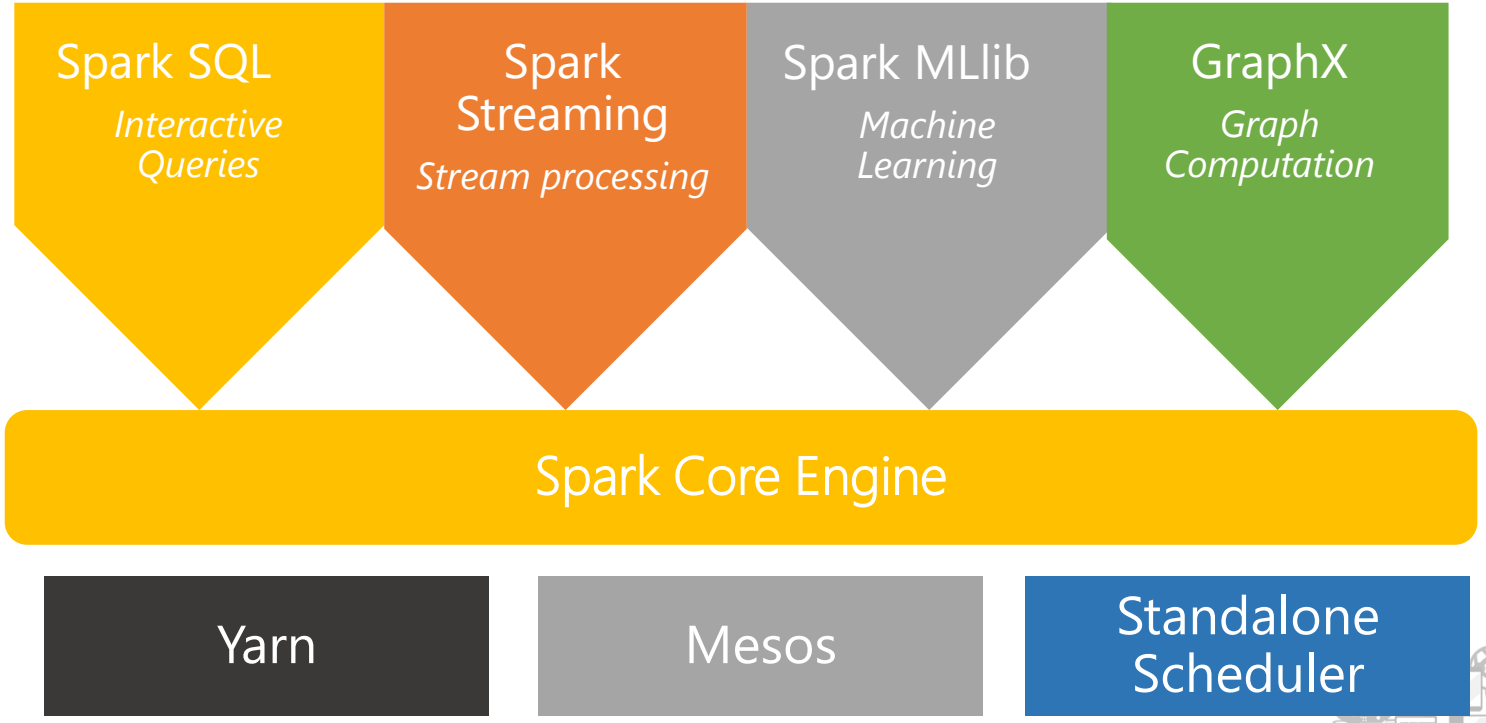


Apache Spark – An Unified Framework

An unified, open source, parallel, data processing framework for Big Data Analytics

Spark Unifies:

- Batch Processing
- Real-time processing
- Stream Analytics
- Machine Learning
- Interactive SQL



Spark - Benefits

Performance

Using in-memory computing, Spark is considerably faster than Hadoop (100x in some tests).
Can be used for batch and real-time data processing.

Developer Productivity

Easy-to-use APIs for processing large datasets.
Includes 100+ operators for transforming.

Unified Engine

Integrated framework includes higher-level libraries for interactive SQL queries, processing streaming data, machine learning and graph processing.
A single application can combine all types of processing

Ecosystem

Spark has built-in support for many data sources such as HDFS, RDBMS, S3, Apache Hive, Cassandra and MongoDB.

Runs on top the Apache YARN resource manager.

Spark – Use cases

Use case	Description	Users
Data Integration and ETL	Cleansing and combining data from diverse sources	Palantir: Data analytics platform
Interactive analytics	Gain insight from massive data sets tin ad hoc investigations or regularly planned dashboards.	Goldman Sachs: Analytics platform Huawei: Query platform in the telecom sector.
High performance batch computation	Run complex algorithms against large scale data	Novartis: Genomic Research MyFitnessPal: Process food data
Machine Learning	Predict outcomes to make decisions based on input data	Alibaba: Marketplace Analysis Spotify: Music Recommendation
Real-time stream processing	Capturing and processing data continuously with low latency and high reliability	Netflix: Recommendation Engine British Gas: Connected Homes

Spark is fast

Spark is the current (2014) Sort Benchmark winner.
3x faster than 2013 winner (Hadoop).

	2013 Record (Hadoop)	Spark 100 TB	Spark 1 PB
Data Size	102.5 TB	100 TB	1000 TB
Time	72 min	23 min	234 min
Nodes	2100	206	190
Cores	50400	6592	6080
Rate/Node	0.67 GB/min	20.7 GB/min	22.5 GB/min

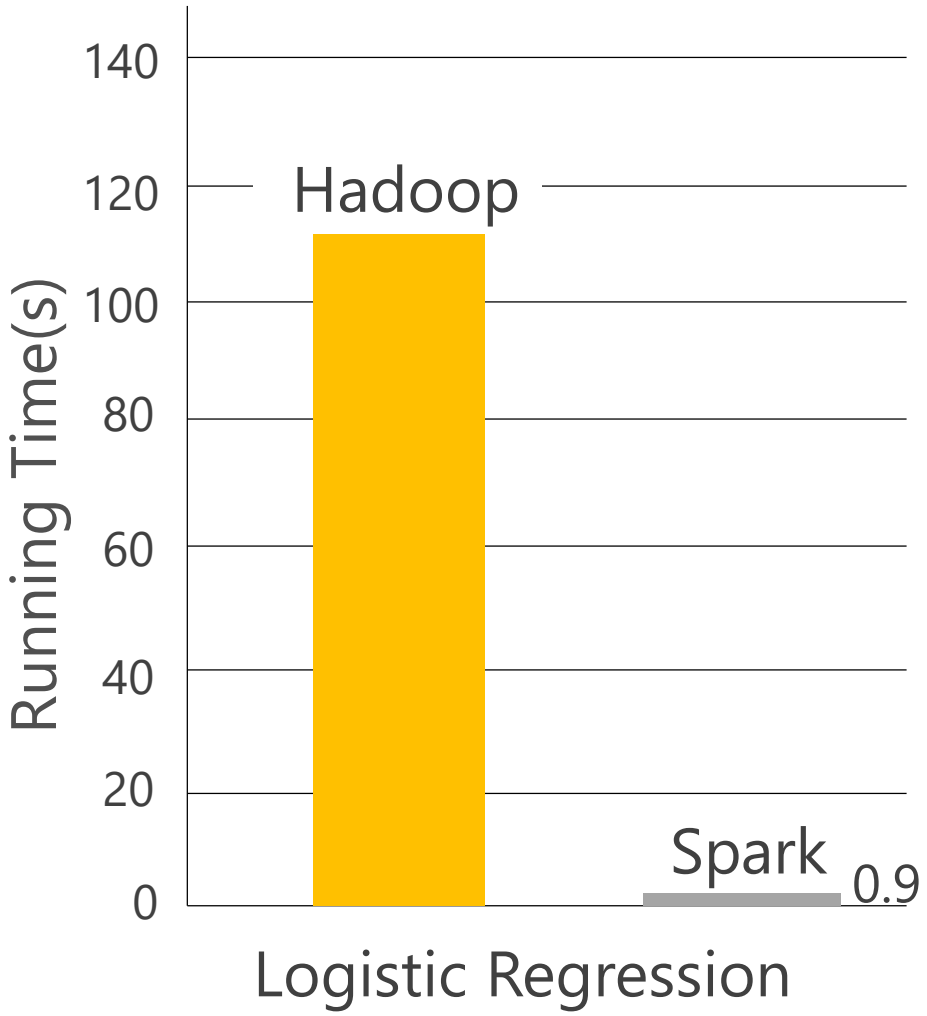
Spark is fast not just for In-Memory but On-Disk computation as well

tinyurl.com/spark-sort

… especially for iterative applications

In iterative applications the same data is accessed repeatedly often in a sequence.

Most machine learning algorithms and streaming applications (that maintain aggregate) state are iterative in nature.



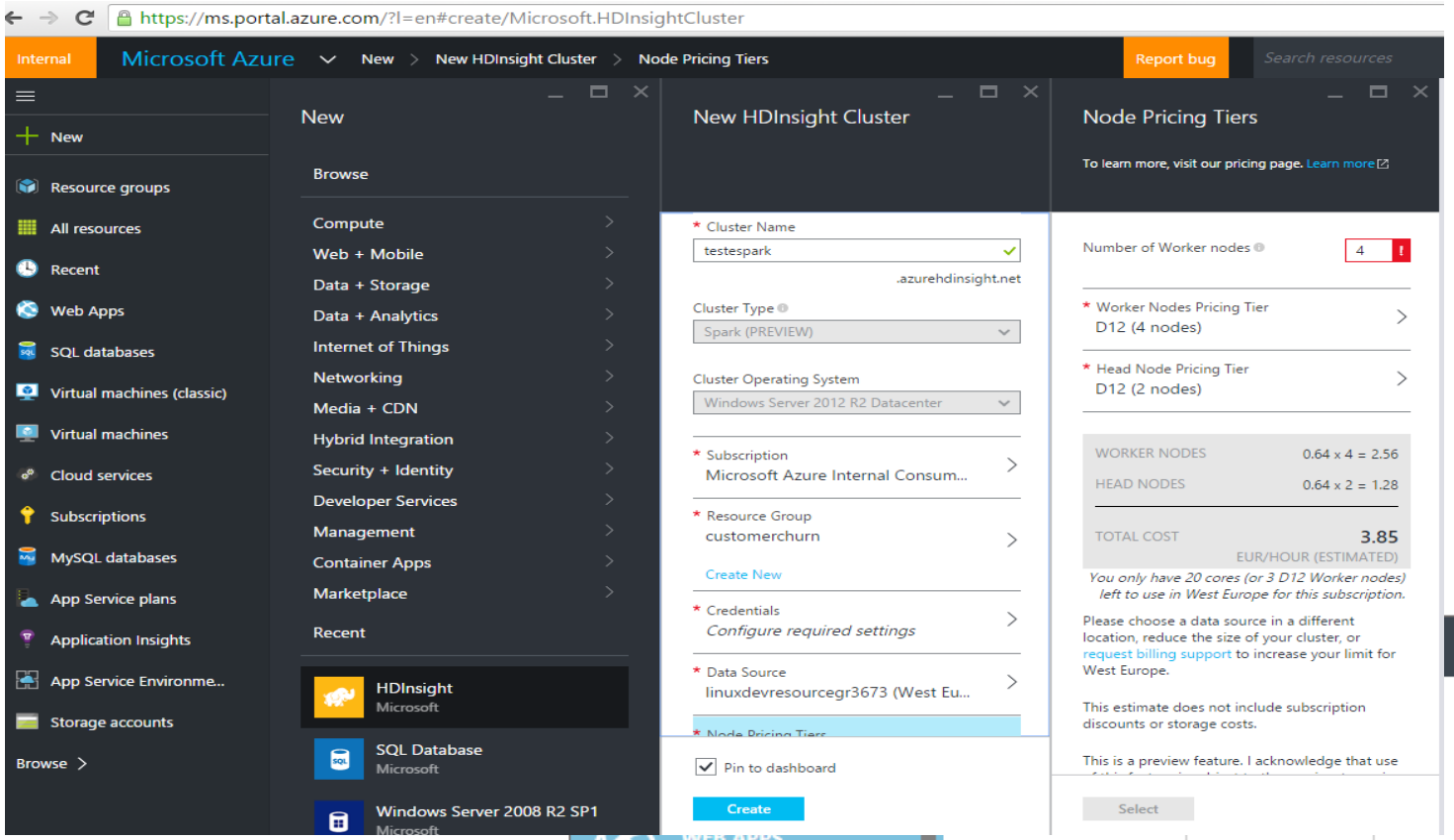
Logistic regression on a 100-node cluster with 100 GB of data

Creating Spark Cluster on Azure

20 mins lead time

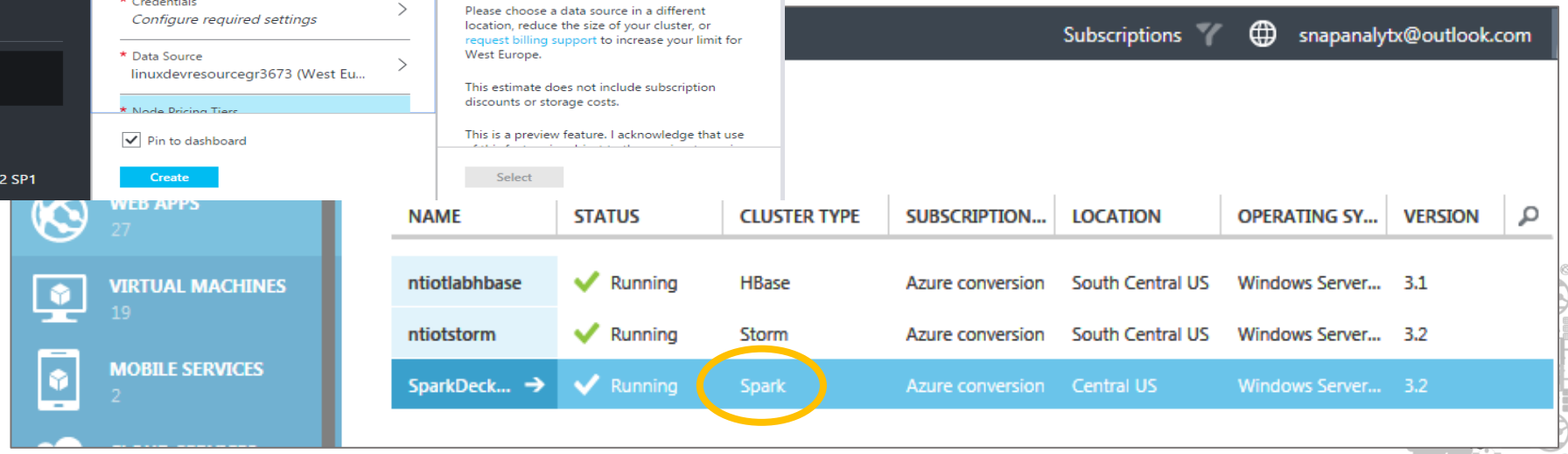
- Time for Coffee Break!?

Creating a HDInsight Spark Cluster



- A Spark cluster can be provisioned directly from the Azure console.
- Only the number of data nodes have to be specified (can be changed later)
- More nodes enable more queries to be run concurrently

The Azure console lists all types of HDInsight clusters (HBase, Storm, Spark etc) currently provisioned



HDInsight Spark Dashboard

Microsoft Azure HDInsight Spark Dashboard

Spark UI | Resource Manager | Notebooks | File Browser | Quick Links | Help + Feedback

Spark 1.3.1 **Spark Master at spark://headnodehost:7077**

URL: spark://headnodehost:7077
REST URL: spark://headnodehost:6066 *(cluster mode)*
Workers: 2
Cores: 8 Total, 2 Used
Memory: 12.0 GB Total, 2.0 GB Used
Applications: 2 Running, 3 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers

Worker Id	Address	State	Cores	Memory
worker-20150703034026-workernode0.SparkDeckSmall.g7.internal.cloudapp.net-49901	workernode0.SparkDeckSmall.g7.internal.cloudapp.net:49901	ALIVE	4 (1 Used)	6.0 GB (1024.0 MB Used)
worker-20150703034027-workernode1.SparkDeckSmall.g7.internal.cloudapp.net-49727	workernode1.SparkDeckSmall.g7.internal.cloudapp.net:49727	ALIVE	4 (1 Used)	6.0 GB (1024.0 MB Used)

Running Applications

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
app-20150703070533-0004	Zeppelin	1	1024.0 MB	2015/07/03 07:05:33	hdp	RUNNING	1.4 h
app-20150703034203-0000	SparkSQL::headnode0	1	1024.0 MB	2015/07/03 03:42:03	hdp	RUNNING	4.8 h

Completed Applications

Application ID	Name	Cores	Memory per Node	Submitted Time	User	State	Duration
app-20150703055747-0002	pyspark	1	1024.0 MB	2015/07/03 05:57:47	hdp	FINISHED	16 min
app-20150703060138-0003	pyspark	1	1024.0 MB	2015/07/03 06:01:38	hdp	FINISHED	9.8 min
app-20150703054926-0001	pyspark	1	1024.0 MB	2015/07/03 05:49:26	hdp	FINISHED	4.7 min

- The HDInsight Spark Dashboard provides links to access:
  - The **Spark UI** (the web-based UI to monitor a Spark cluster)
  - The **Resource Manager** that controls the amount of resources allocated to various Spark cluster components
  - **Notebooks** – interactive web based tools to develop and run Spark programs

HDInsight Spark Resource Manager

Microsoft Azure HDInsight Spark Dashboard

Spark UI | Resource Manager | Notebooks | File Browser | Quick Links | Help + Feedback

Resource Manager

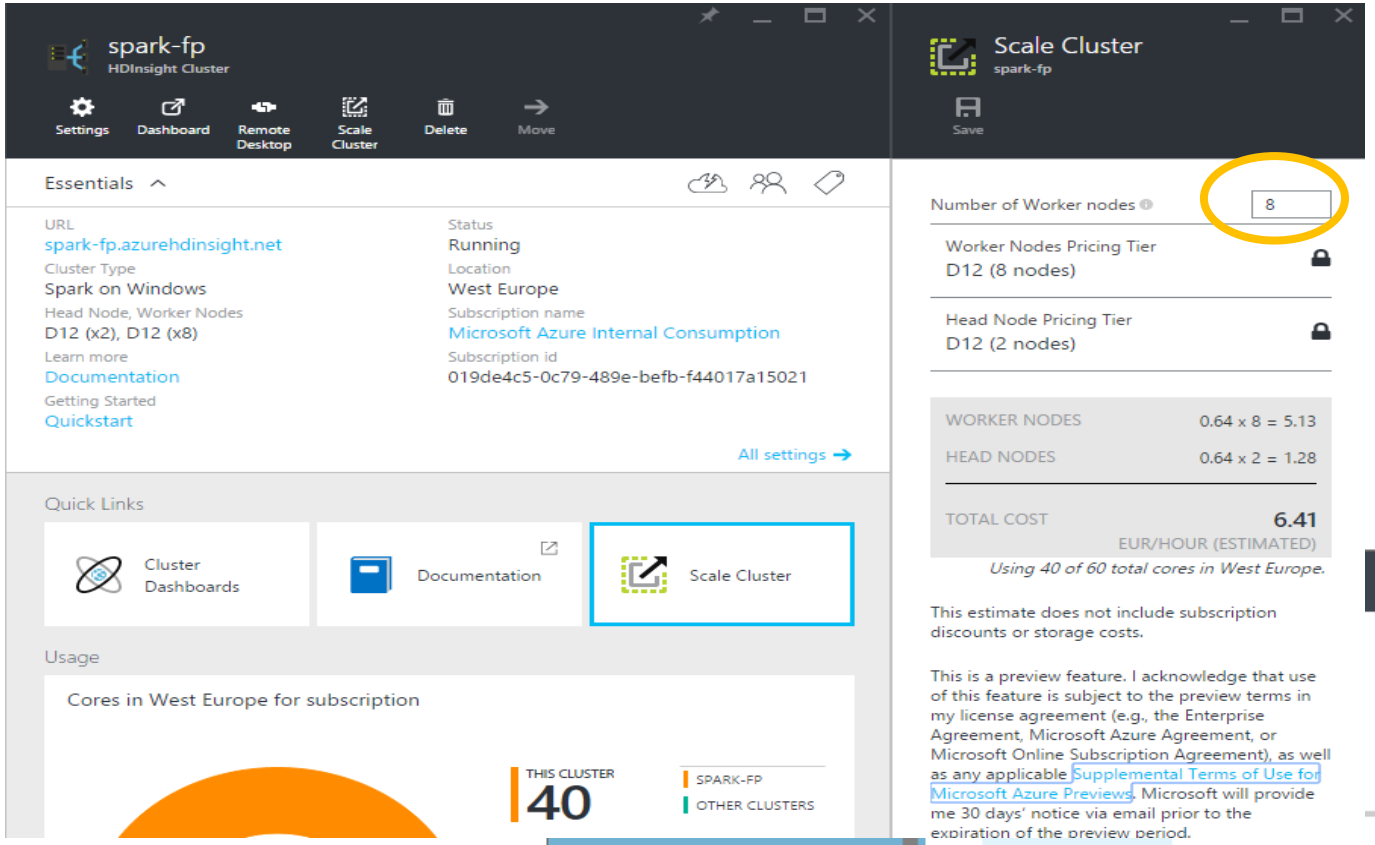
Application Setting	Property	Value
Default application core count [Spark-defaults.conf]	spark.cores.max	2
Core count [Thrift Server]	spark.cores.max	2
Core count [Zeppelin]	spark.cores.max	2
Default executor memory per worker node [Spark-defaults.conf]	spark.executor.memory	1g
Executor memory per worker node [Thrift Server]	spark.executor.memory	1g
Executor memory per worker node [Zeppelin]	spark.executor.memory	1g

Restore default values | Submit

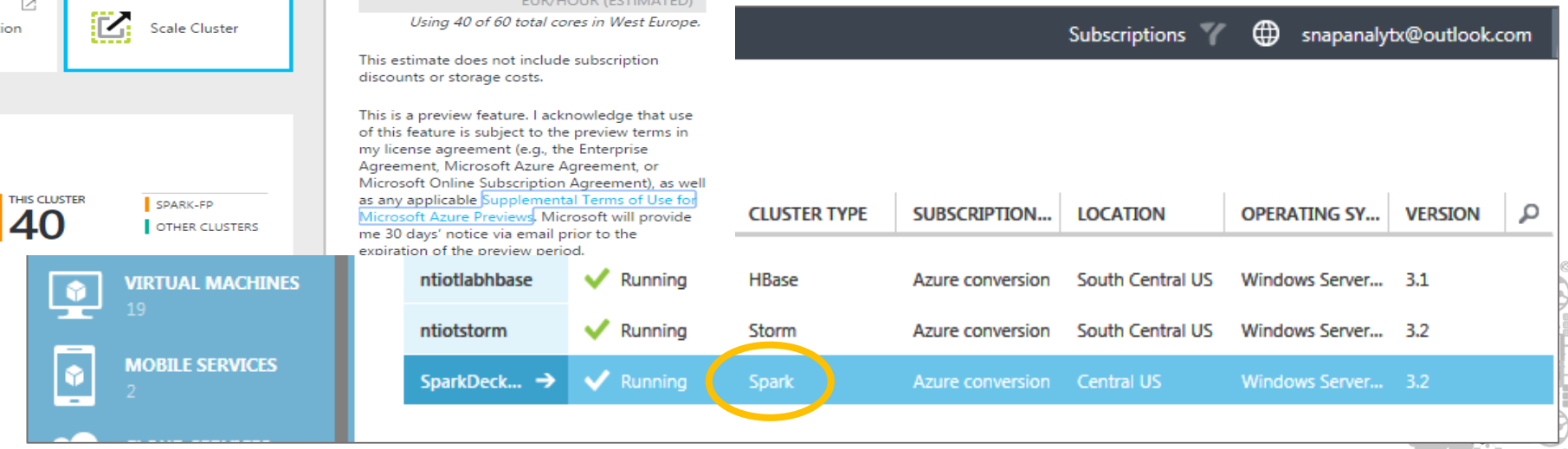
The Resource Manager enables you to control the number of cores and amount of memory allocated to Spark cluster components and notebooks.

Increasing the resources allocated to the Thrift Server can potentially improve the performance with BI Tools

Resizing a HDInsight Spark Cluster



- Like other HDInsight cluster types, the number of nodes in a Spark cluster can be increased dynamically
- The cluster or the running jobs do not have to be paused or stopped to resize
- Note: Performance of queries can also be improved by caching data in memory or SSD



Developing with Notebooks

Developing Spark Apps with Notebooks

Notebooks:

- Are web-based, interactive servers for REPL (Read-Evalute-Print-Loop) style programming.
- Are well-suited for prototyping, rapid development, exploration, discovery and iterative development
- Typically consist of code, data, visualization, comments and notes
- Enable collaboration with team members

Jupyter and Zeppelin are two Notebooks that work with Apache Spark

Interactive Queries with SparkSQL

Interactive Analytics

- With PowerBI and Tableau

Spark SQL Overview

- An extension to Spark for processing *structured data*.
  - Part of the core distribution since Spark 1.0 (April 2014)
- Is a distributed SQL query engine
- Supports SQL and HiveQL as query languages
- Also a *general purpose* distributed data processing API.
- Binding in Python, Scala and Java
- Can query data stored in external databases, structured data files (eg JSON), Hive tables etc more. [See spark packages for a full list of sources that are currently available]

Integration with BI Reporting Tools

HDInsight Spark integrates with these BI tools to report on Spark data

Microsoft Power BI

tableau

QlikView

SAP Lumira

Machine Learning with Spark MLlib

A Recommender System Example

- Movie Recommendation Model

What is MLlib?

- A collection of machine learning algorithms optimized to run in a parallel, distributed manner on Spark clusters for better performance on large datasets
- Seamlessly integrates with other Spark components
- MLlib applications can be developed in Java, Scala or Python

Type	Algorithms
Supervised	Classification and Regression: • Linear Models (SVMs) logistic regression, linear regression) • Naïve Bayes • Decision Trees • Ensembles of trees (Random Forest, Gradient-Boosted Trees) • Isotonic regression
Unsupervised	Clustering: • k-means and streaming k-means • Gaussian mixture • Power iteration clustering (PIC) • Latent Dirichlet allocation (LDA)
Recommendation	Collaborative Filtering • Alternating least squares (ALS)

Movie Recommendation – Dataset

- Will use the publicly available "*MovieLens 100*k" dataset.
- It is a set of 100,000 data points related to ratings given by users to a set of movies
  - It also includes movie metadata and user profiles. (not needed for recommendation)
- The dataset can be downloaded from http://files.grouplens.org/dataset

User Ratings Data (u.data)

User Id	Movie Id	User's Rating	Timestamp
196	242	3	881732314
198	302	3	883894932
22	377	1	883443433
145	51	2	886570342
187	356	4	885634452
166	63	5	886554545

- Each user has rated several movies and at least one movie
- The ratings vary from 1 to 5
- The fields in the file (u.data) are tab separated.
- Users and movies are identified by Id. More details about the movie and profile of the users are in the other files (u.item and u.user) respectively.

Go Try Yourself

Links

- https://azure.microsoft.com/pt-pt/pricing/free-trial/
- https://azure.microsoft.com/en-us/documentation/articles/hdinsight-apache-spark-overview/
- https://azure.microsoft.com/pt-pt/services/hdinsight/apache-spark/
- https://azure.microsoft.com/pt-pt/documentation/articles/hdinsight-apache-spark-ipython-notebook-machine-learning/
- https://azure.microsoft.com/pt-pt/documentation/articles/hdinsight-apache-spark-zeppelin-notebook-jupyter-spark-sql/

Videos

- https://channel9.msdn.com/Shows/Azure-Friday/Announcing-Apache-Spark-on-Azure-HDInsight
- https://channel9.msdn.com/Shows/Data-Exposed/Data-Analysis-with-Tableau-and-Spark-on-Azure-HDInsight
- https://channel9.msdn.com/Series/Azure-Data-Lake/Whats-up-with-Spark15-Spark-Architecture

500+

New releases in the last 12 months

Azure Site Recovery: Protect VMWare and Physical Servers in Public Preview

Azure Backup Generally Available

Azure API Management Premium simplifies high availability and massive scale for APIs

ExpressRoute for Office 365

Azure Active Directory Dynamic Membership For Groups

Automatic Password Change for Social Media Shared Accounts

Compute-Intensive A10 and A11 Virtual Machine Instances

Remote Desktop app for Windows Phone support for Gateway and Remote Resources

Informatica Cloud Agent availability in Linux and Windows Virtual Machines

Azure DocumentDB Hadoop Connector

Azure HDInsight support for more VM sizes

Enterprise-Grade Array-Based Replication and Disaster Recovery

Free Azure Trial

http://aka.ms/tryazure

Try SQL Server 2016 CTP2

http://aka.ms/trysql2016

Use Power BI for Free

http://powerbi.microsoft.com