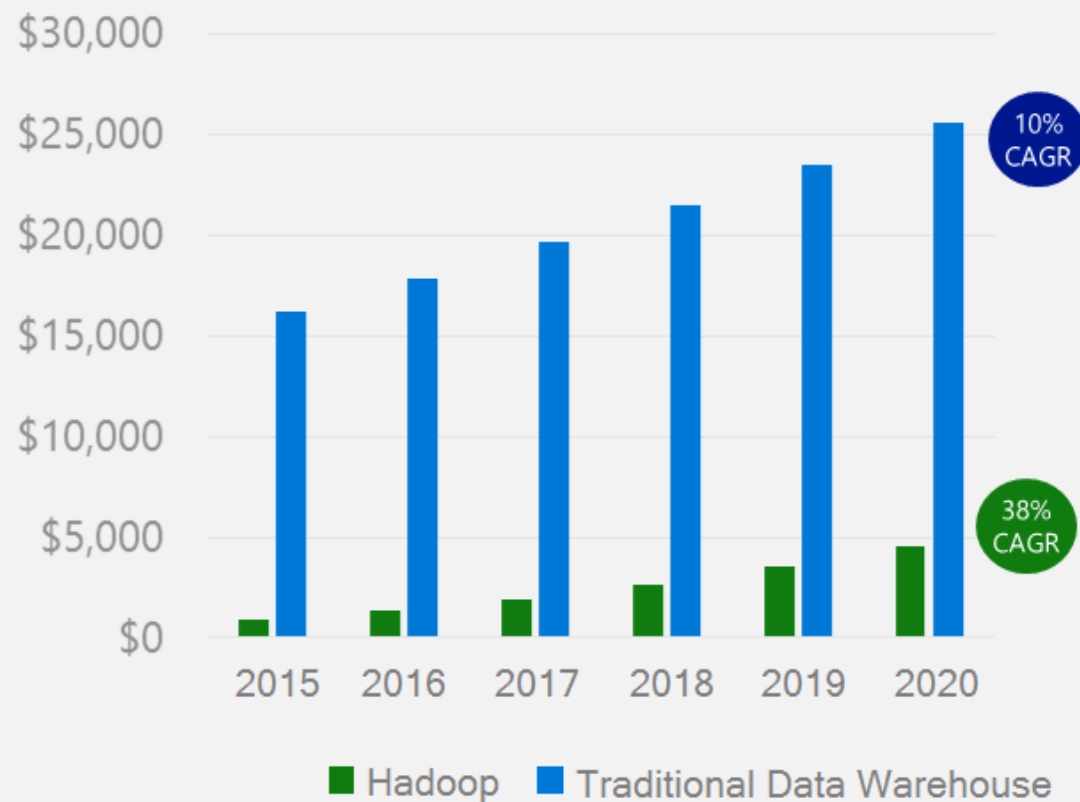


微软大数据及数据仓库概述

Speaker Name

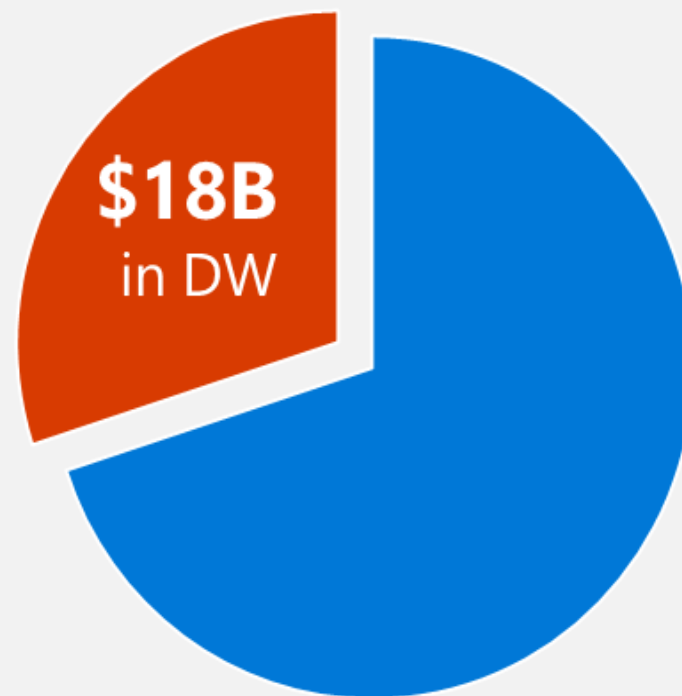
大数据及数据仓库的巨大商机

Data Warehousing and Hadoop
Market Revenues



*451 Research Total Data Market Monitor Q2 2016

DW represents 30% of the
\$60B data platform opportunity



微软云高级分析服务解决方案

解决方案

HDInsight/Spark



大数据平台

大规模并行计算 企业级规模，一次写入，
分布式部署

数据科学家/
数据工程师

Microsoft R



基于R的分析

数据科学家

Azure 机器学习



云分析

便于建模，分析
及对比

数据科学家

认知服务

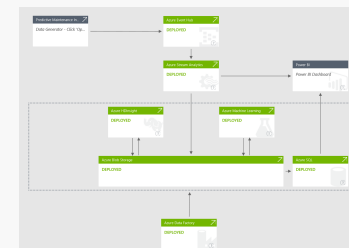


分析API服务

便捷的视觉、语音、语言、
知识等的分析

开发者

解决方案



预配置的应用及
解决方案

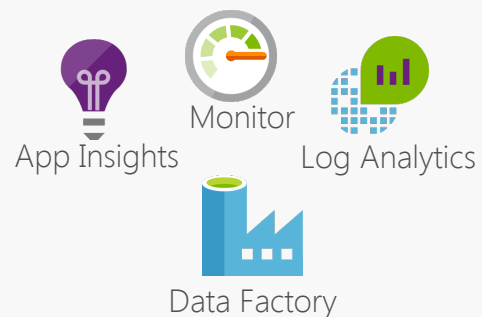
解决特定行业/业务/领域
的应用/解决方案

BDM/TDM

大数据负载及相关服务

大数据典型场景及工作负载

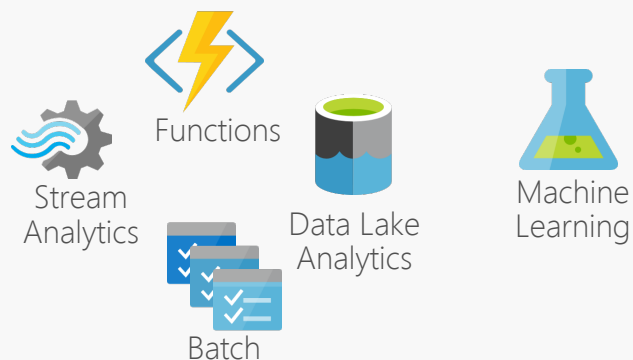
数据获取/ETL



流处理



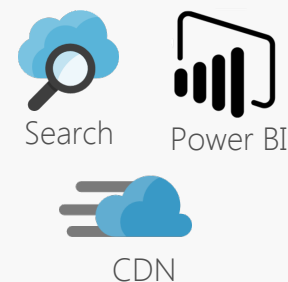
分析及机器学习



数据聚合



数据展现



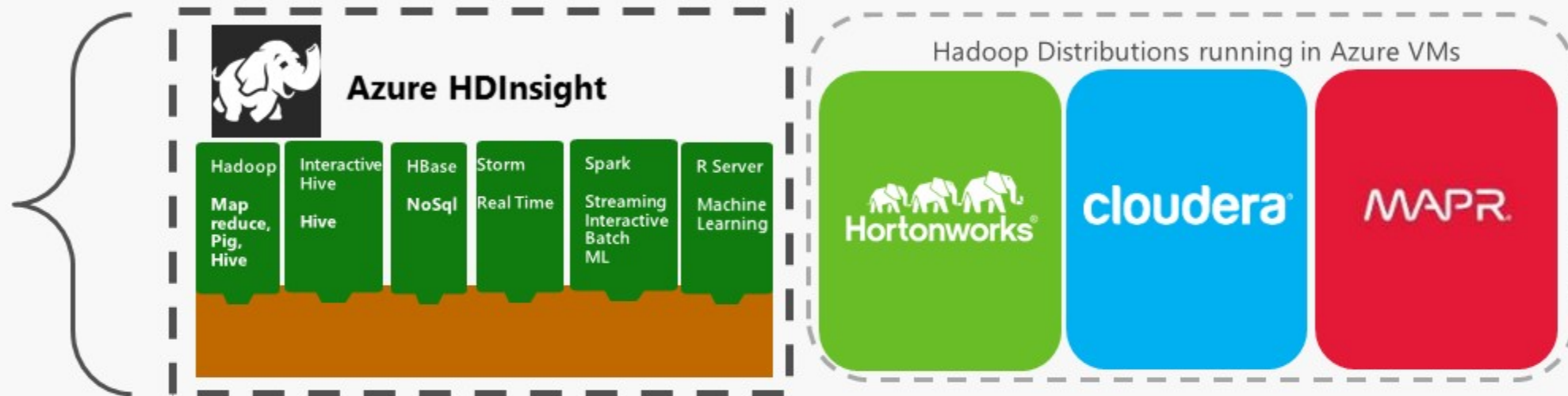
Azure HDInsight提供的相关服务



Azure 智能云基础设施

Microsoft Hadoop 技术栈

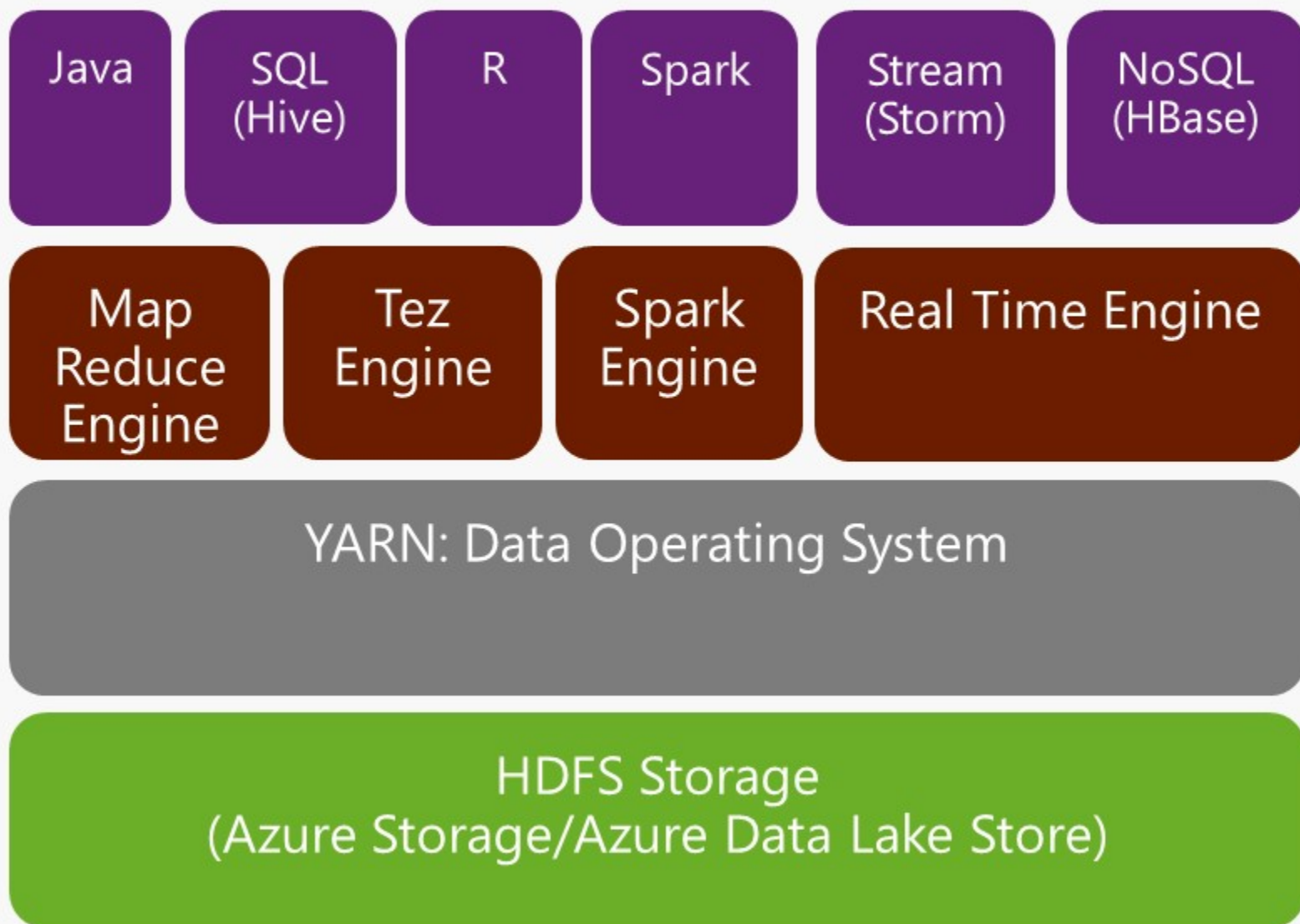
分析



存储

Local (HDFS) or Cloud (Azure Blob/Azure Data Lake Store)

HDInsight 开放的技术架构



Microsoft R 技术

- Microsoft R Open
- Microsoft R Client
- Microsoft R Server...
...for Hadoop, ...for Teradata, ...for Linux (SUSE, Red Hat/CentOS)
- Microsoft SQL Server 2016/2017 R Services
- Microsoft R Open in Azure ML
- Microsoft R Scripting in PowerBI

Microsoft R Server

R+CRAN

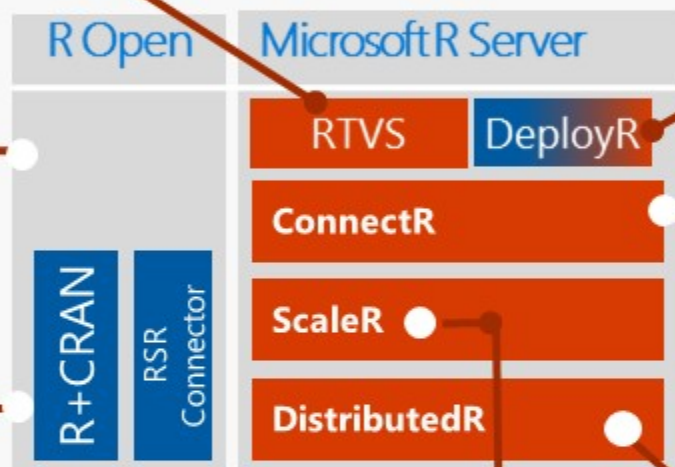
- Open source R interpreter
 - R 3.1.2
- Freely-available huge range of R algorithms
- Algorithms callable by RevoR
- Embeddable in R scripts
- 100% Compatible with existing R scripts, functions and packages

Microsoft R Open

- Based on open source R
- High-performance math library to speed up linear algebra functions
- Checkpoint package to easily share R code and replicate results using specific R package versions

R Tools for Visual Studio

- State of the art, R Tools for Visual Studio IDE



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

DeployR

- RESTful APIs for easy integration from Java, JavaScript, .NET
- Enterprise authentication & security
- Horizontal scaling

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

基于HDI Insight提供高伸缩性的机器学习平台



Cluster configuration

* Cluster Type ⓘ
R Server ▼

* Operating System
Linux

* Version
R Server 9.0 (HDI 3.5) ▼

* Cluster Tier ⓘ
STANDARD PREMIUM

R Server : Terabyte-scale, enterprise grade R analytics with transparent parallelization on top of Spark and Hadoop.

Configuration Options:

- R Server 9.0.0 on Spark 2.0.0 with Java 8
- R Server 8.0.5 on Spark 1.6.2 with Java 7

Adds \$0.02 per Core-Hour.

Features

* denotes preview feature

Available	Not available
<input checked="" type="checkbox"/> R Studio community edition for R Server	+ Apache Ranger* (PREMIUM) ⓘ
+ Secure shell (SSH) access	+ Domain joining* (PREMIUM) ⓘ
+ HDInsight applications	+ Remote Desktop access ⓘ
+ Custom virtual network	
+ Custom Hive metastore	
+ Custom Oozie metastore	
+ Data Lake Store access	



Cluster configuration

* Cluster Type ⓘ
Spark ▼

* Operating System
Linux

* Version
Spark 2.0.1 (HDI 3.5) ▼

* Cluster Tier ⓘ
STANDARD PREMIUM

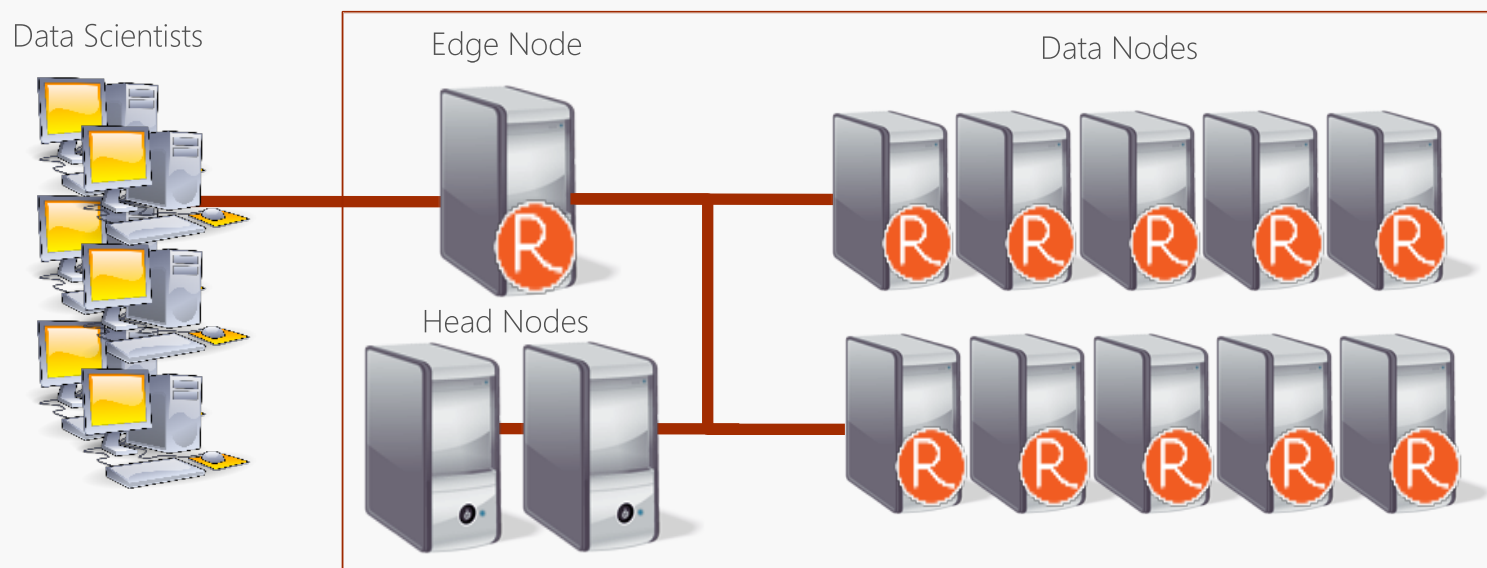
Spark : Fast data analytics and cluster computing using in-memory processing.

Features

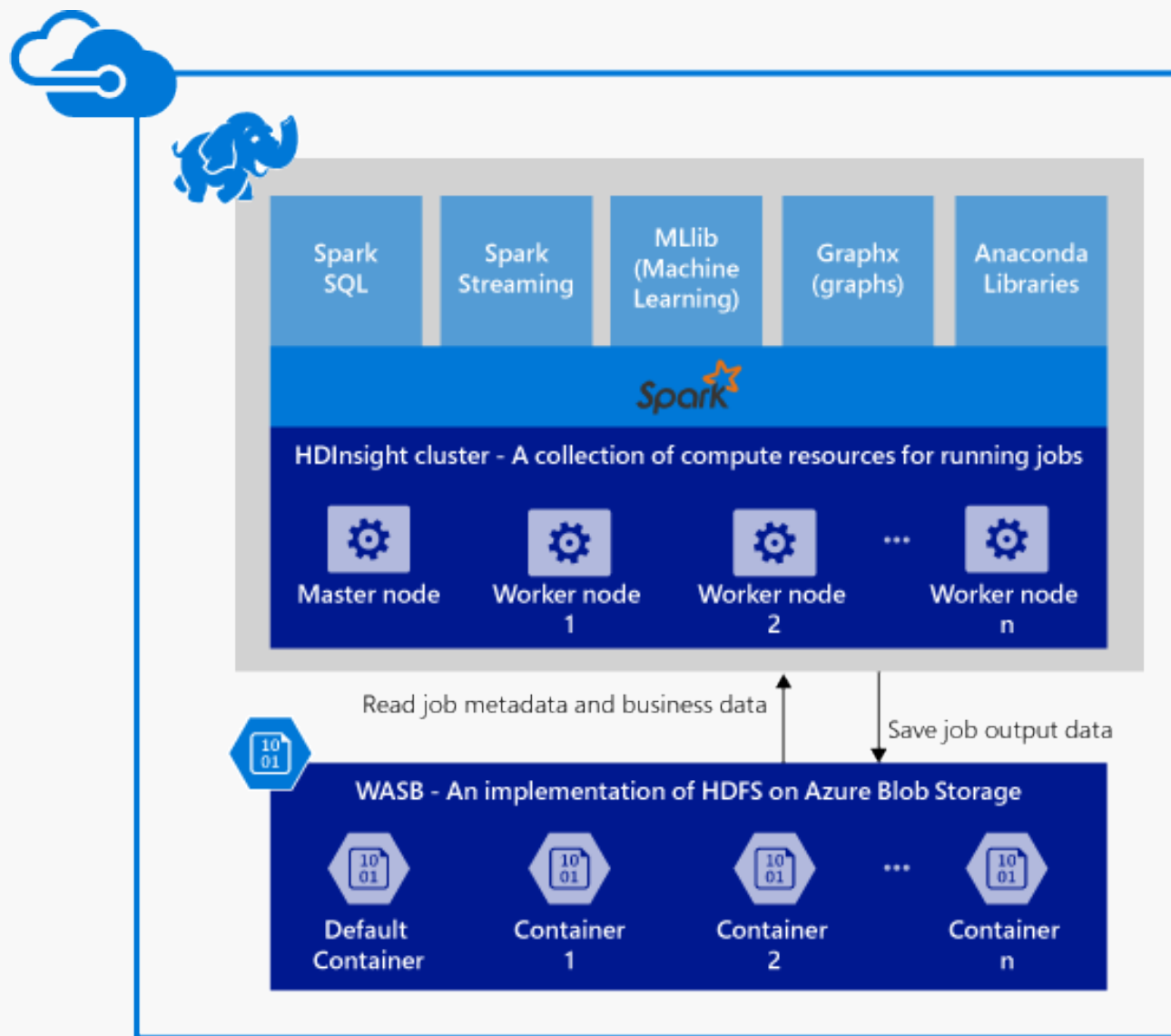
* denotes preview feature

Available	Not available
+ Secure shell (SSH) access	+ Apache Ranger* (PREMIUM) ⓘ
+ HDInsight applications	+ Domain joining* (PREMIUM) ⓘ
+ Custom virtual network	+ Remote Desktop access ⓘ
+ Custom Hive metastore	
+ Custom Oozie metastore	
+ Data Lake Store access	

HDI Insight中的R Server 集群及快速伸缩

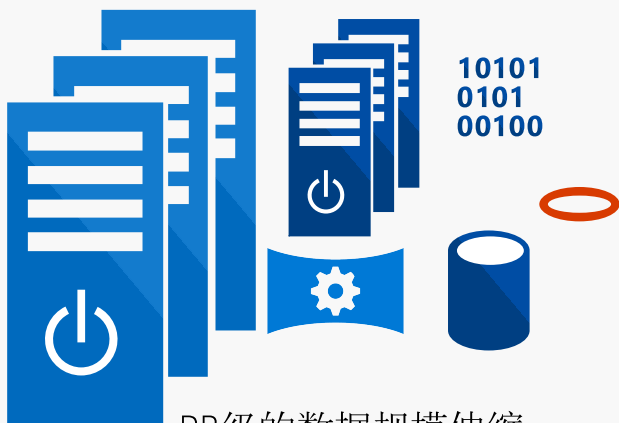


Azure HDInsight中的Spark集群



Azure SQL 数据仓库

弹性伸缩及性能

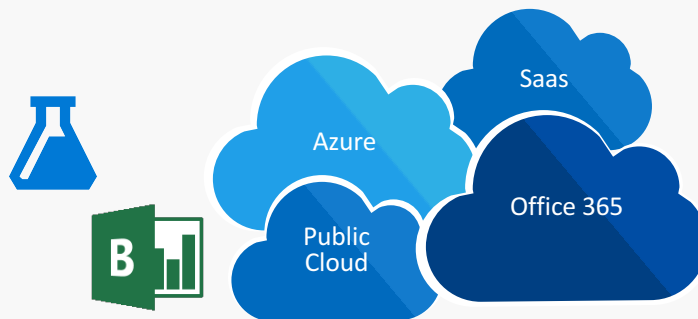


PB级的数据规模伸缩
大规模并发处理（MPP）
秒级的快速伸缩
结构化/非结构化查询

基于云的强大支撑

快速部署及获得

与Azure ML, PowerBI及ADF的紧密集成

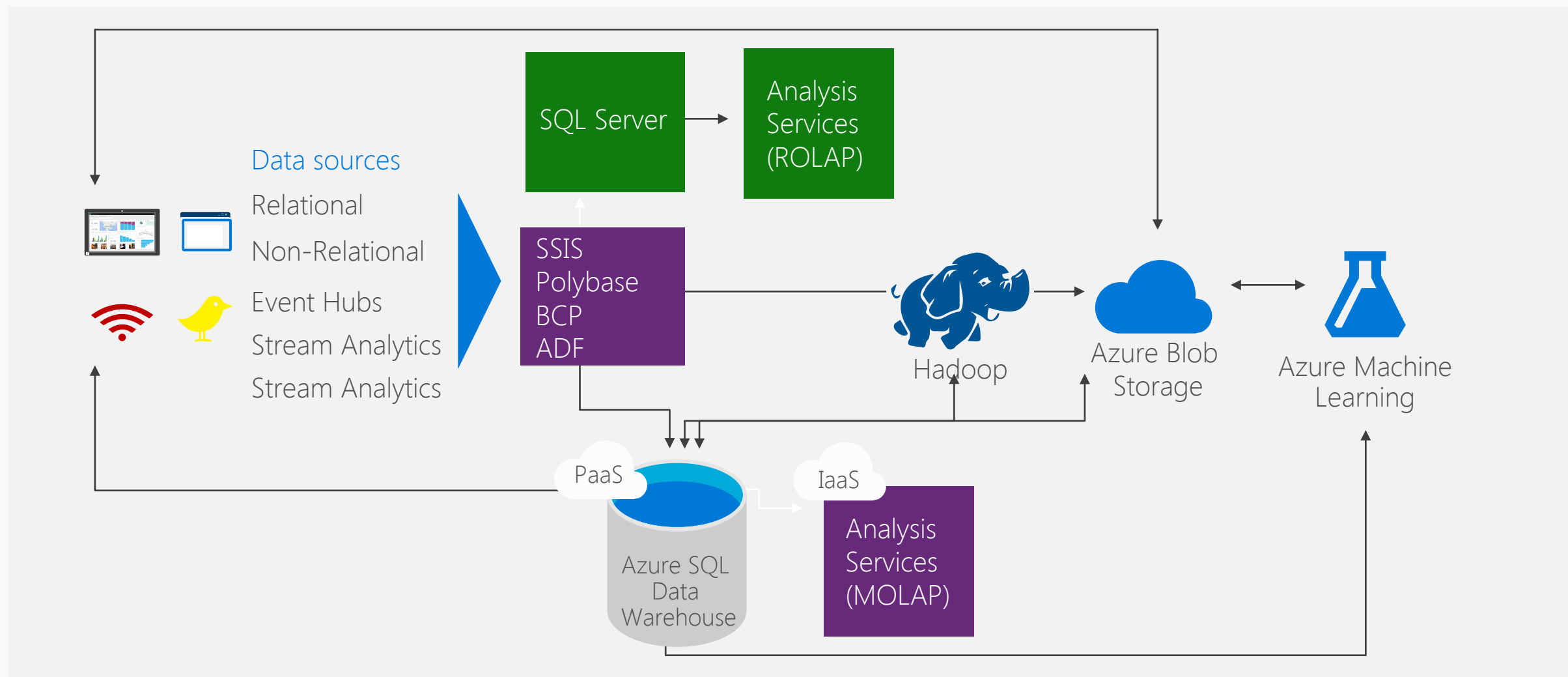


领先的成本及性能优势



计算及存储的简便支付及获得
按需付费，无需等待

Azure SQL 数据仓库，全面支持高级分析需要



谢谢！