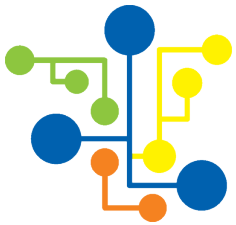

Data Science Academy
Analytics - Visualização, Relatórios e Tomada de Decisões com Big Data

Big Data Analytics com Azure HDInsight – Parte 1

Analytics, Visualização, Relatórios e Tomada de Decisões com Big Data



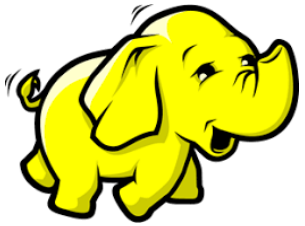
Data Science Academy

Big Data Analytics com Azure HDInsight Parte 1

- Hadoop e HDInsight
- Tipos de Clusters, Tipos de Nodes
- Criação de Clusters
- Consultas com Hive
- Acesso a máquinas do Cluster via ssh
- Criação de Banco de Dados no Azure
- Exportação de dados com sqoop

Hadoop

- Criado em 2005 por Mike Cafarella e Doug Cutting
- Faz parte da Apache Software Foundation
- Plataforma de código-aberto para o armazenamento e processamento distribuído de grandes conjuntos de dados, utilizando clusters de computadores
- Projetado para ser escalável a milhares de máquinas, fazendo uso do processamento e armazenamento de cada uma
- Tolerante a falhas, entregando aplicações com alta-disponibilidade.
- Dados não-estruturados
- Armazenamento, processamento, acesso, governança, segurança, etc



Por que utilizar o Hadoop na nuvem?

- Software livre X software gratuito
- Pago pelo uso (computação e armazenamento de dados) => Economia
- Escalonável
- Seguro e em conformidade

A nuvem é flexível e oferece dimensionamento rápido

Na nuvem do Microsoft Azure, você paga somente pela computação e o armazenamento usados. Crie um cluster Hadoop, analise seus dados e desligue-o para interromper o medidor.

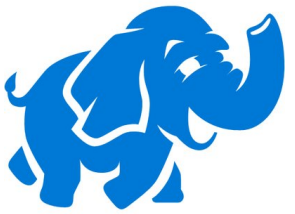
NHS

Criamos rapidamente o cluster do Azure HDInsight e processamos seis anos de dados em apenas algumas horas, depois desativamos&ellipsis. Processar os dados na nuvem torna o processo muito acessível.

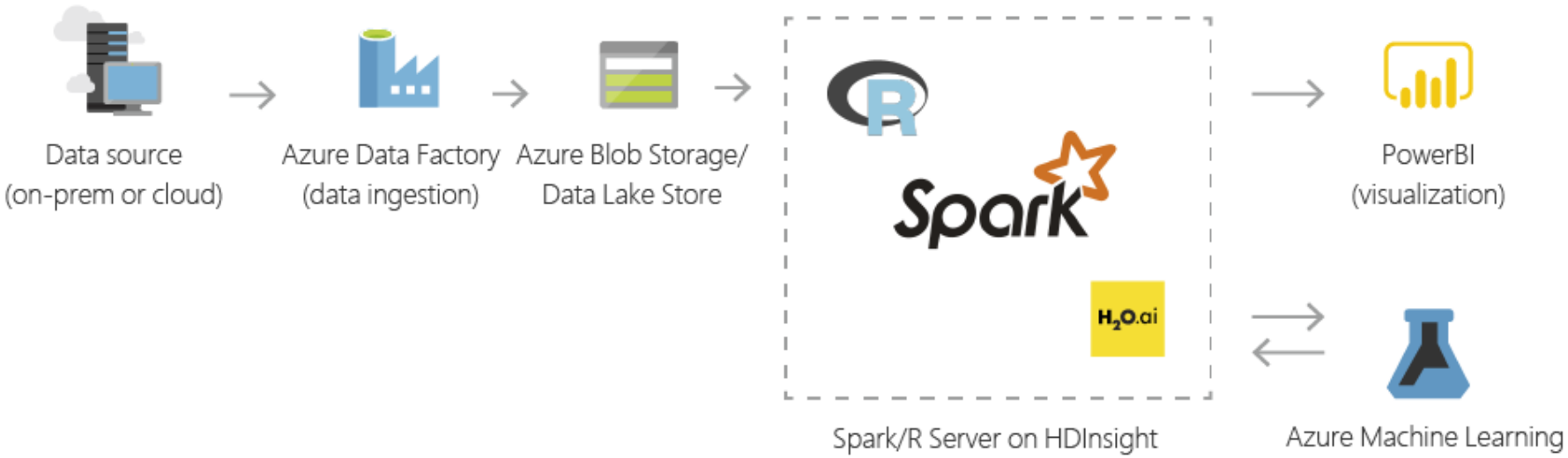
–Paul Henderson, Serviço Nacional de Saúde do Reino Unido

Microsoft Azure HDInsight

- Serviço em nuvem totalmente gerenciado para processamento de grandes volumes de dados
- Softwares livres: Hadoop, Spark, Hive, Kafka, Storm, R, etc
- Cenários: ETL, Data Warehousing, Machine Learning, IoT, etc
- Criação de clusters sob demanda
- Computação e armazenamento desacoplados

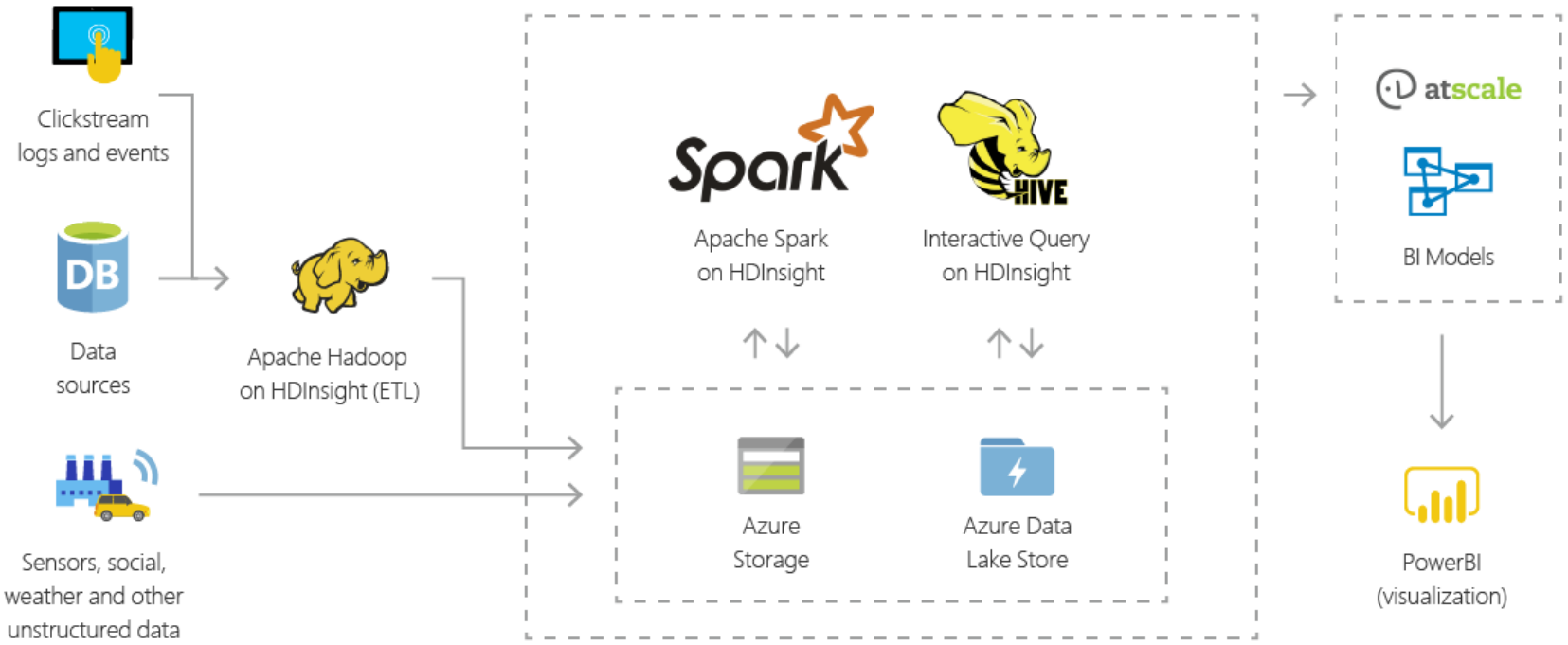


Cenário 1: Ciência de Dados + ML



Caso de uso: https://customers.microsoft.com/en-us/story/pros

Cenário 2: Data Warehouse



Tipos de Cluster Hadoop

- Apache Hadoop: utiliza HDFS, gerenciamento de recursos YARN e modelo de programação MapReduce para processar e analisar dados em lote de forma paralela.

- Apache Spark: fornece suporte ao processamento em memória para melhorar o desempenho dos aplicativos de análise de Big Data.

- Apache Hbase: banco de dados NoSQL baseado no Hadoop que fornece acesso rápido para grandes quantidades de dados desestruturados e semi-estruturados, potencialmente com bilhões de linhas vezes milhões de colunas.

Tipos de Cluster Hadoop

- R Server: fornece a cientistas de dados, estatísticos eprogramadores em R, acesso sob demanda a métodos escalonáveis e distribuídos de análise de dados no HDInsight.

- Apache Storm: sistema de computação distribuído e em tempo real para processar rapidamente grandes fluxos de dados.

- Visualização de Consulta do Apache Interactive Hive: cache na memória para consultas Hive interativas e rápidas.

- Apache Kafka: fornece funcionalidade de fila de mensagem, para criação de aplicativos e pipelines de transmissão de dados

Tipos de Nós

- Head Nodes: coordena o processo de execução dos jobs entre os demais nodes. Possui 2 nós principais.

- Worker node: realizam o processamento

- Edge node: disponível no cluster R Server, permite testar o código R localmente no nó antes de enviá-lo ao cluster para processamento distribuído.

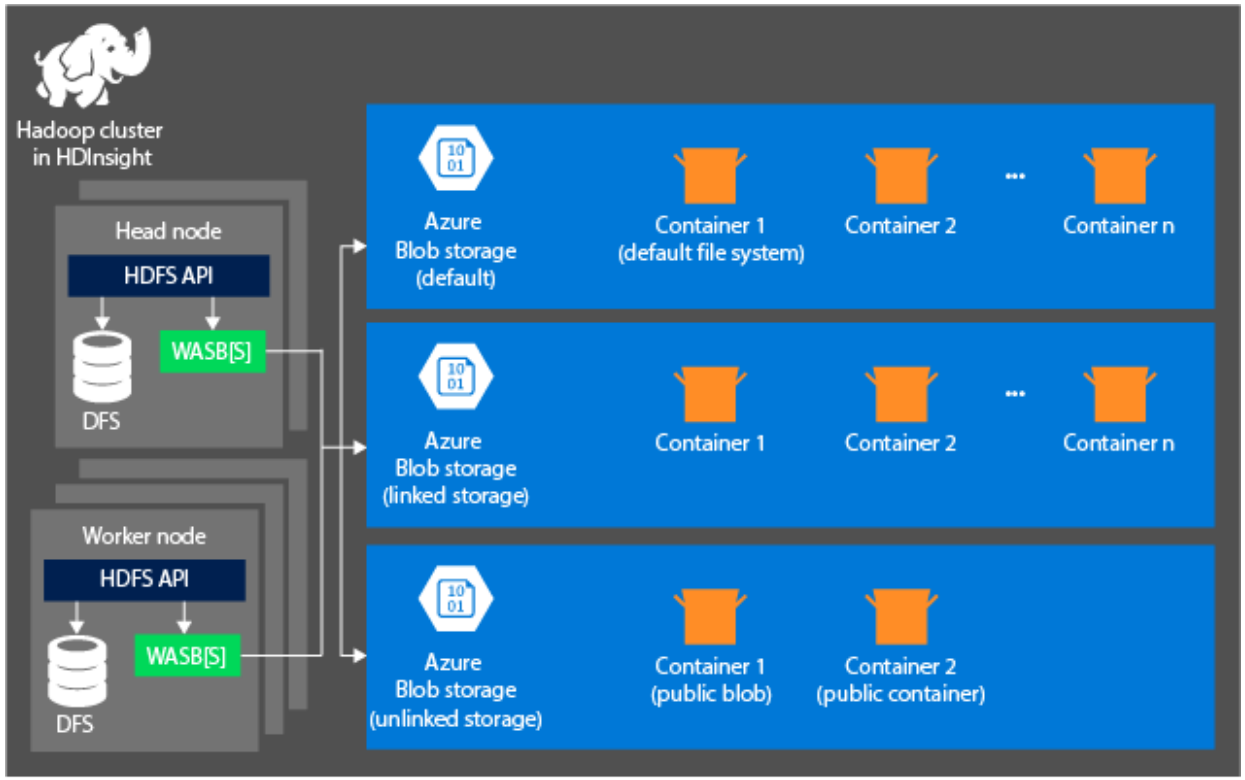
Linguagens de Programação

- Compatíveis com Java e Python
- Suporte a Scala, Clojure, Jython (Python para Java)
- Linguagens do ecossistema Hadoop: Pig Latin, HiveQL

Opções de Armazenamento para clusters HDInsight

- Azure Storage
- Azure Data Lake Store
- Ambos

Opção de armazenamento Azure Storage



Acesso ao sistema de arquivos local: hdfs://<nomedohost>/<caminho>
Acesso ao sistema de armazenamento do azure: wasb[s]://<nomedocontainer>@<nomedaconta>.blob.core.windows.net/<path>

Opção de armazenamento Data Lake Store

adl://meudatalake/<cluster_root_path>

<cluster_root_path> é o nome da pasta raiz criada para cada cluster no Data Lake Store

adl://meudatalake/clusters/cluster_storage_1

adl://meudatalake/clusters/cluster_storage_2

Opção de armazenamento - Benefícios

- Compartilhamento e reutilização de dados
- Arquivamento de dados
- Custo
- Escala horizontal elástica
- Replicação geográfica
- Data Lake Analytics (USQL, etc)

Tipos de Instâncias e preços

https://docs.microsoft.com/pt-br/azure/cloud-services/cloud-services-sizes-specs#size-tables

https://azure.microsoft.com/pt-br/pricing/details/hdinsight

Apache Hive

- Sistema de DW para Hadoop
- Permite a projeção de estrutura em grandes volumes de dados não-estruturados
- Consultas Hive são escritas em HiveQL, semelhante ao SQL. Não é necessário conhecimento de Java ou do MapReduce
- Manual da linguagem HiveQL: https://cwiki.apache.org/confluence/display/Hive/LanguageManual

Tabelas Hive

- Internas: dados são armazenados no Data Warehouse do Hive, localizado no armazenamento padrão do cluster no caminho /hive/warehouse

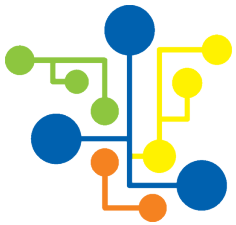
- Externas: dados são armazenados fora do Data Warehouse. Os dados podem estar armazenados em qualquer armazenamento acessível pelo cluster.

Planejamento de Clusters

- Região
- Local de armazenamento
- Tamanho do armazenamento
- Tipo de Cluster
- Capacidades das máquinas virtuais
- Dimensionamento do Cluster
- Ciclo de vida do Cluster
- Cotas

FIM

Analytics, Visualização, Relatórios e Tomada de Decisões com Big Data



Data Science Academy