



Curso 03: Administração de Dados Complexos em Larga Escala -- Comparação Hadoop x Spark --

Prof. Jose Fernando Rodrigues Junior

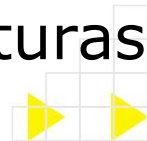
Objetivo: comparar as soluções de processamento Big Data Hadoop e Spark



Níveis da Análise de Dados



- **Análise de dados básica:** contagens, somas, médias, máximo, mínimo, e ordenação;
- **Análise de dados estatística:** distribuição de dados, ajuste de modelo, teste de hipóteses, métricas, etc;
- **Análise de dados avançada:** aprendizado de máquina, classificação, regressão, recomendação, clusterização, etc;
- **Aprendizado de máquina avançado:** arquiteturas de redes neurais visando inteligência artificial.



Níveis da Análise de Dados

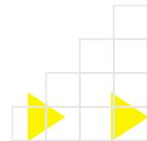


- **Análise de dados básica:** médias, máximo, mínimo, e ordem; Curso 03/11 - DW/OLAP
- **Análise de dados estatística:** distribuição de dados, ajuste de modelo, t-testes, métricas, etc; Curso 02/05
- **Análise de dados avançada:** aprendizado de máquina, classificação, regressão, clusterização, etc; Curso 02/03 - Spark
- **Aprendizado de máquina avançado:** arquiteturas de redes neurais visando inteligência; Curso 05/07/08/10

Comparativo



vs



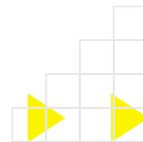
Hadoop



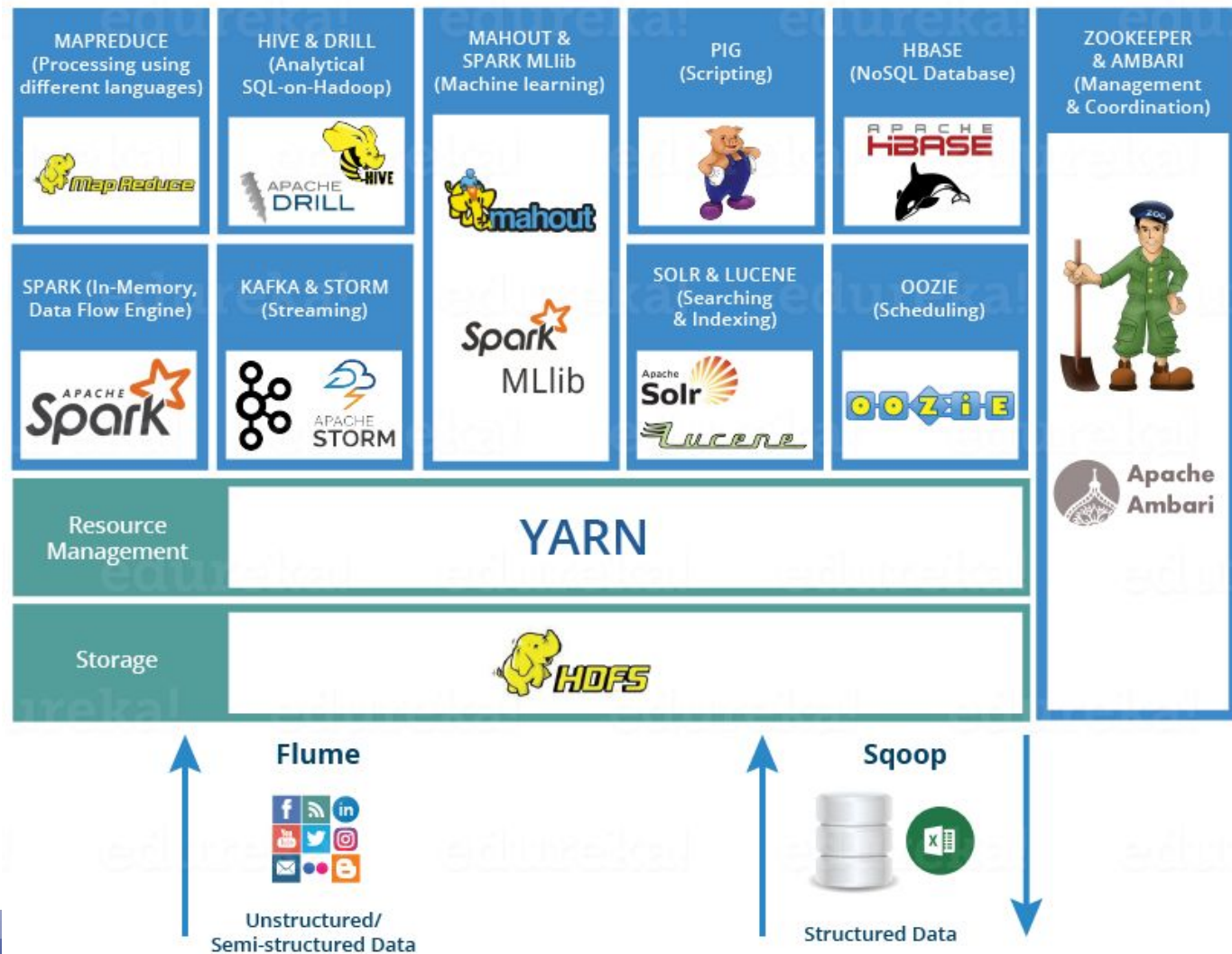
- O Apache Hadoop é um **arcabouço de código aberto** para armazenar e processar com eficiência conjuntos de dados que variam de gigabytes a petabytes;
- um projeto que ganhou força já a partir de 2005;
- Sua escala é conhecida por abranger **Petabytes de dados** - Facebook, Yahoo, Expedia, British Airways,....;
- Dezenas de outros softwares **trabalham em conjunto** ou sobre a combinação HDFS+MapReduce: Hive, HBase, Kylin, Mahout, Pig, etc

⇒ [The Hadoop Ecosystem Table](#)

⇒ [Hadoop at Amazon AWS](#)



É tudo sobre
Big Data.

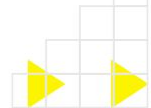


Hadoop



⇒ Limitações:

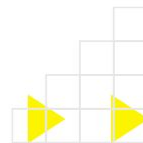
- não é a melhor solução para **poucos dados** - alguns Terabytes por exemplo;
- não funciona de modo **interativo**, mas em processamento batch - inadequado para *real-time processing*;
- difícil integração com dados produzidos de maneira contínua (***streaming***);
- processamento **iterativo** dificultado, ou inviável;
- uso mais **complexo** do que Spark - Java apenas, e projeto complexo.



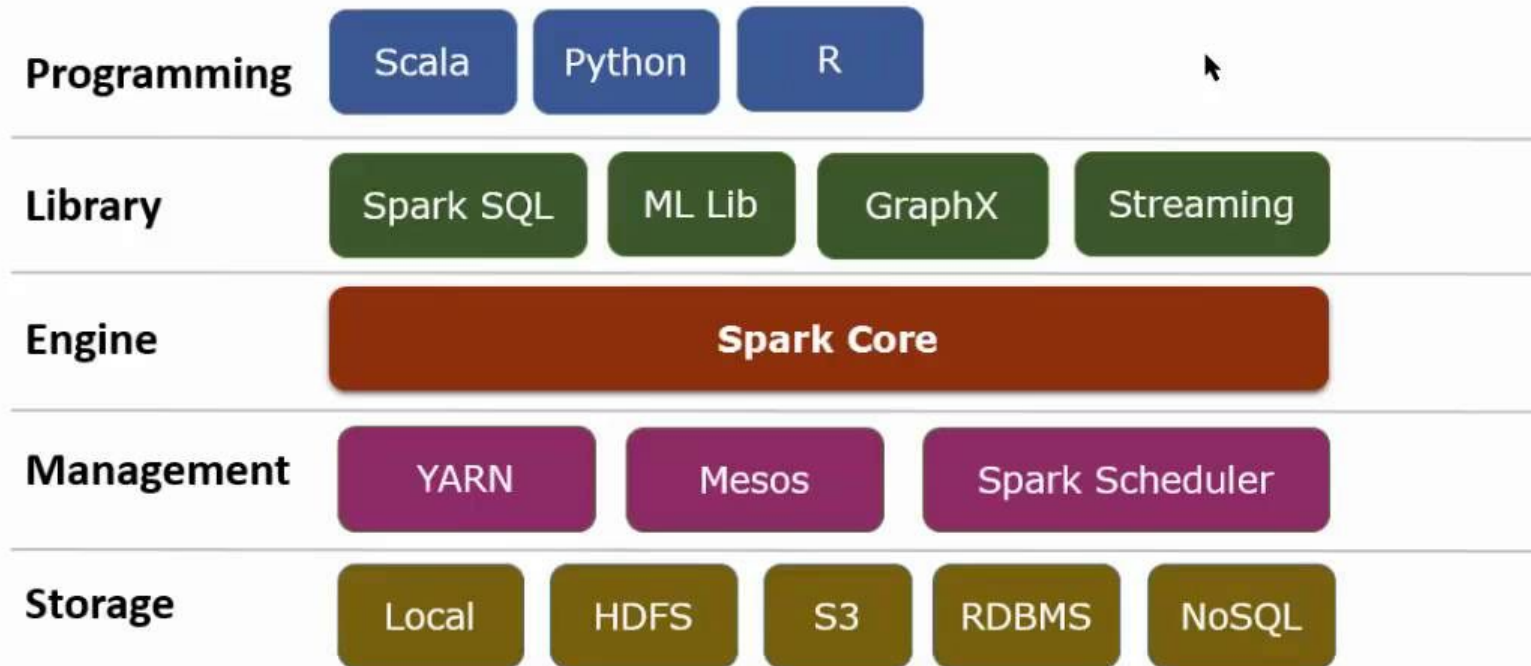
Spark



- Arcabouço de **processamento distribuído**;
- Faz uso de processamento em memória, ao invés de fluxo sequencial de dados - 10 a 100 vezes **mais rápido do que Hadoop**;
- Consegue **sanar as limitações** do Hadoop:
 - . linguagens de programação: Python, Scala, Java, e R;
 - . lida bem com dados em streaming;
 - . faz processamento iterativo ou batch.
- **Integrável** com diversos subsistemas: HDFS, HBase, Yarn, Mesos, Cassandra, ou standalone.



Spark Framework



Comparativo



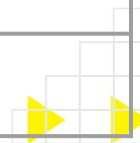
| Hadoop | Spark |
|--|--|
| Processamento +Armazenamento distribuídos | Processamento distribuído |
| MapReduce | Resilient Distributed Dataset (RDD) |
| Disk-intensive | Memory-intensive |
| Batch | Iterativo (grafos, proces. numérico) e Interativo (real-time, spark shell); e também batch |
| Java (embora existam outras iniciativas) | Scala, Python, Java, e R (simplicidade de uso) |



Comparativo



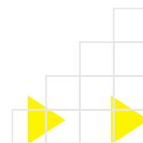
| <i>Task</i> | Hadoop | Spark |
|-------------------------------|------------------|----------------------------------|
| Data Warehouse | Hive | Spark SQL |
| ETL | Sqoop, Flume | Spark SQL |
| Aprendizado de máquina | Mahout | Machine Learning Library (MLlib) |
| Streaming | Storm e/ou Kafka | Spark Streaming |
| Real-time analytics | None | Suporta |
| Grafos | Inviável | GraphX |



Spark sobre o Hadoop



- Embora versátil, o Spark surgiu no contexto do **ecossistema Hadoop** - uma evolução;
- Aceita outras integrações, mas, comumente, faz uso do **HDFS e do YARN**;
- Hadoop = HDFS+MapReduce, o que torna o conceito “usa Hadoop” **nebuloso**;
- O Spark não é um substituto para o Hadoop por duas razões:
 - usa parte de seu legado;
 - ainda que uma evolução, sua escala é menor.

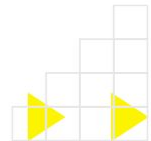


Spark



⇒ Limitações

- requer que **memória disponível** nos nós de processamento - se tiver que fazer paginação (*virtual memory*), o desempenho cai drasticamente;
- a necessidade de **memória eleva o custo**; por exemplo, se deseja-se um cluster com máquinas de baixo custo, o Hadoop é uma solução mais adequada.



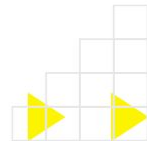
Conclusões



- **Hadoop**, muito bom em:
 - processamento simples;
 - altíssima escala;
 - sobre clusters de baixo custo;
- **Spark**, muito bom em:
 - processamento simples ou complexo;
 - alta escala;
 - sobre clusters de custo mais elevado.

⇒ Critérios:

- qual a escala dos meus dados?
- quais computadores eu já tenho?
- quais problemas eu preciso resolver?



Conclusões



ORACLE®



MySQL™



No SQL Databases



Key Value



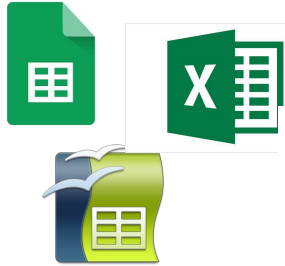
Document Db



Wide Column



Graph DB



algumas
linhas

KBytes

MBytes

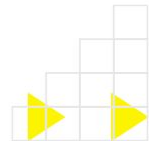
GBytes

Terabytes

Centenas de
TBytes

Petabytes

Centenas de
Petabytes



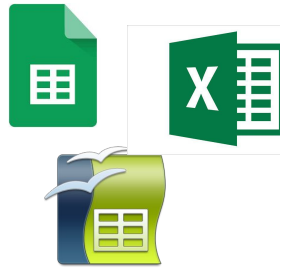
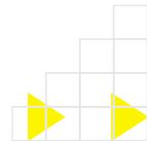
Conclusões



loop



s de
es



algumas
linhas

KBy

INFRASTRUCTURE

The collage displays logos for various cloud services categorized into four groups:

- STORAGE:** Includes logos for Amazon S3, Google Cloud Storage, Microsoft Azure Storage, IBM Cloud Object Storage, Oracle Cloud Infrastructure, and Pure Storage.
- HADOOP CLOUDERA:** Includes logos for Hadoop, Cloudera, and other related ecosystem components.
- DATA LAKES:** Includes logos for AWS Lake Formation, IBM Data Platform, and other services designed for managing large-scale data lakes.
- DATA WAREHOUSES:** Includes logos for Amazon Redshift, Google BigQuery, Microsoft Azure Synapse Analytics, Snowflake, and others.
- STREAMING / IN-MEMORY:** Includes logos for Apache Kafka, Amazon Kinesis, and other real-time data processing services.

| NO-SQL DATABASES | NEW-SQL DATABASES | GRAPH DBs | MPP DBs | SERVER LESS | CLUSTER SVCS |
|---|--|--|--|--|--|
|  Google Cloud SQL  Amazon Redshift  ORACLE  MarkLogic  Databricks  Microsoft Azure  ScyllaDB |  Snowflake  Amazon Aurora  Couchbase  MongoDB  Microsoft SQL Server  IBM Db2  AragoDB |  Neo4j  Amazon Neptune  ORACLE  Hiboot  Microsoft SQL Server  IBM Db2 |  Teradata  VERTICA  Celonis  Sapientino  EcoSQL |  Amazon S3  Google Cloud Storage  IBM Cloud  Microsoft Azure  Google Cloud Storage  IBM Cloud |  Amazon ElastiCache  IBM Cloud  Microsoft Azure  Google Cloud Storage  IBM Cloud  Microsoft Azure |

| ETL / DATA TRANSFORMATION | DATA INTEGRATION | DATA GOVERNANCE | DATA QUALITY |
|---|---|---|---|
|  Talend  Pentaho  Alteryx  Informatica Data Integration  Microsoft Data Integration  DataStage  IBM DataStage  Paxata  StreamSets |  Data Services  Sequester  MuleSoft  Informatica  Anaplan  Stitch  Data Lake Catalog  Troy.io  Fivetran  AT&T  Zalando  Import.io  MATillion  InfoWorks  Snowplow  Narrator  Census  Greenleaf |  Informatica Data Governance  DataPoint  IBM Data Governance  Alation  Collibra  Mulesoft Data Governance  Dremio  IBM Data Governance  Okera  DataHub  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus |  Talend Data Quality  TORO  SODA  DataSift  DataHub  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus  DataNexus |

| | | | | |
|---|--|---|---|---|
| MGMT / MONITORING                | DATA GENERATION & LABELING       | AI OPS ALGORITHMS        | CPU DBs & CLOUD       | AI HARDWARE                |
|---|--|---|---|---|

ANALYTICS & MACHINE INTELLIGENCE

The image displays a collection of logos for various data science and analytics platforms, organized into four main categories:

- BI PLATFORMS:** Includes logos for Looker, Tableau, Microsoft Power BI, Qlik, Alteryx, and others.
- VISUALIZATION:** Includes logos for Tableau, Microsoft Power BI, Qlik, Alteryx, and others.
- DATA ANALYST PLATFORMS:** Includes logos for Microsoft, Alteryx, Tableau, and others.
- DATA SCIENCE PLATFORMS:** Includes logos for Microsoft, Alteryx, Tableau, and others.

The diagram illustrates the relationship between Data Science, Data Science Platforms, and Machine Learning. On the left, a large blue circle labeled "DATA SCIENCE" contains three sub-sections: "DATA SCIENCE NOTEBOOKS" (listing JupyterLab, Colab, Google Cloud DataLab, Kaggle, Orange3, and Neo4j), "DATA SCIENCE PLATFORMS" (listing Binder, Microsoft Cloud, Oracle Data Science Platform, SAS, TIBCO, KNIME, Domino, Alteryx, and RapidMiner), and "MACHINE LEARNING" (listing Microsoft, H2O.ai, DataRobot, Gamalan, VIZEN, Element8, deepstream.ai, and Octolabs). Arrows point from the Data Science Platforms section to the Machine Learning section, indicating a flow or relationship between the two.

[illegible]

SEARCH

- elasticsearch
- amazon
- ORACLE
- EMERSON
- PRESTIGE
- Google Analytics
- algolia
- covivio
- SINEQUA
- Lucidworks
- ATTIVIO
- swiftype
- EXALEAD
- alphasense
- emrnlv
- MAANA

LOG ANALYTICS

- splunk
- Google
- Microsoft
- Sumologic
- elastic
- loggly
- logz
- timber
- loggly
- logz

SOCIAL ANALYTICS

- Netbase
- hootsuite
- springr
- synthesio
- simplereach
- bitly
- Twitter
- socialWith

WEB / MOBILE / COMMERCE ANALYTICS

- Google Analytics
- piwikpro
- SIGOPT
- Airtable
- RESCI
- granify
- Amplitude

– APPLICATIONS – ENTERPRISE

| SALES | MARKETING - B2B | MARKETING - B2C | CUSTOMER EXPERIENCE / SERVICE | HUMAN CAPITAL |
|---|---|---|---|---|
|  arista  chorus  conversica  aviso  peoplea  cloudi  tactical  si  machines |  App Annie  Lattice  sense  Relativity  tubular  ENGAGIO |  VEEVA  ACTION  Segment  Simon  theother  Compucon  Sendbird  CONTENT SQUARE  zelus  Sparkroll  Amperio  Bloomreach  Bluecore  INVOCAC  buzago [PERASO] | quintus  SurveyMonkey  Capterra Testing  CLARABRIDGE  MEDALLIA  zendesk  Customer  freshdesk  Gainsight  pendo  HEAP  Amplitude  Veeva Analytics  Delighted  USERCENTRIC  INTERCOM  Duoft ASAPP  afiniti  netomi  Catalix  aerotel  Eureka  talia  Home.ai |  HireVue  Symmetry  TECTIA  mya  Alyio  Wade & Wenderly  E |

[illegible]

APPLICATIONS – INDUSTRY

| ADVERTISING | EDUCATION | REAL ESTATE | GOV'T & INTELLIGENCE | COMMERCE | FINANCE & LENDING | INSURANCE |
|---|---|---|---|--|---|---|
| AppNexus MediaMath Criteo IAS Interactive Oracle Advertising Albert GumGum Oppler The Trade Desk Tafab | Lullaboo VTS 猿辅导 OpenRoads Knewton Orchard eClerx economy360 Skiffree SpaceMach Geophy Koritz | Redfin VTS Opener Orchard economy360 Skiffree SpaceMach Geophy Koritz | Palantir Opener DataMinr MARK42 Anduril FiscatNote Quid PRIMER | FAIR STITCH FIX HowGood STANDARD AYASDI KENSCH Adepara NUMERA | Affirm Monedo TALA ZEST Upgrade Bancor aURA Upstart 100Credit | ROOT Acromia Amenade Shift Technology CAPE EALTEAL Zelus Zesty |

[illegible]

- OPEN SOURCE

The diagram illustrates a comprehensive ecosystem of big data technologies, organized into twelve functional categories:

- FRAMEWORKS**: Includes Hadoop, Spark, Tez, MapReduce, YARN, Flink, and Kubernetes.
- QUERY / DATA FLOW**: Includes Spark SQL, Hive, Pig, Impala, SLAMDATA, and Flink.
- DATA ACCESS & DATABASES**: Includes MongoDB, Redis, Cassandra, Clickhouse, InfluxDB, Druid, Kudu, HBase, Aerospike, and others.
- ADMINISTRATION & PIPELINES**: Includes Talend, Informatica, etcd, Ansible, Puppet, Chef, SaltStack, and others.
- STREAMING & MESSAGING**: Includes Kafka, RabbitMQ, Apache Pulsar, Amazon Kinesis, and others.
- STAT TOOLS & LANGUAGES**: Includes R, Python, Scala, Julia, and others.
- AI OPS & INFRA**: Includes TensorFlow, PyTorch, MXNet, and others.
- AI / MACHINE LEARNING / DEEP LEARNING**: Includes TensorFlow, PyTorch, MXNet, and others.
- SEARCH**: Includes Elasticsearch, Solr, Sphinx, and others.
- LOGGING & MONITORING**: Includes ELK Stack, Splunk, Prometheus, Grafana, and others.
- VISUALIZATION**: Includes Tableau, Power BI, QlikView, and others.
- COLLABORATION**: Includes Confluence, Jira, Slack, and others.
- SECURITY**: Includes Apache Ranger, Knox, and others.

DATA SOURCES & APIs

— DATA RESOURCES

LOCATION INTELLIGENCE

FOURSQUARE mapbox sense360

primary business PlaceIQ esri

CARTO A Radar Mapillary

cubeia OpenStreetMap

OTHER

DATA.GOV

IMxGENET

Lab4Life

bioethics.org

APOLLOSCALE CRUX

DATA SERVICES

QUANTUMBLACK

Booz | Allen | Hamilton

kaggle

ElectrifiAI fractal EXL

DataKind inceptivus

INCUBATORS & SCHOOLS

FULLERLIGHT

GENERAL ASSEMBLY

DataCamp

DataLift

galvanize

INSIGHT

The Data Incubator

RESEARCH

OpenAI

facebook research

MIRI

VECTOR INSTITUTE

AI2