BCC 362 - SISTEMAS DISTRIBUIDOS

VICTOR LOTT
GABRIEL LANA
ARTUR CUNHA

APACHE

Spark™

O QUE É O SPARK?

APACHE SPARK

- Apache Spark é um framework de código fonte aberto para computação distribuída. Foi desenvolvido no AMPLab da Universidade da Califórnia e posteriormente repassado para a Apache Software Foundation que o mantém desde então. Spark provê uma interface para programação de clusters com paralelismo.

EMPRESAS UTILIZANDO O SPARK

- UC Berkeley AMPLab
- 4Quant
- Act Now
- Agile Lab
- Alibaba Taobao
- Alluxio
- Alpine Data Labs
- Amazon
- Art.com
- AsiaInfo
- Atigeo
- atp
- Autodesk
- Baidu
- Bakdata
- Big Industries
- Bizo
- Celtra
- ClearStory Data
- Concur
- Content Square
- Conviva
- Credit Karma
- Databricks
- Dianping.com
- Digby
- Drawbridge
- eBay Inc.
- Elsevier Labs
- EURECOM
- Exabeam
- Faimdata
- Falkonry
- Flytxt
- Freeman Lab at HHMI
- Fundacion CTIC
- GraphFlow, Inc.
- Groupon
- Guavus
- Hitachi Solutions
- The Hive
- IBM Almaden
- InfoObjects
- Inspur
- Istanbul Sehir University
- Kenshoo
- Kelkoo
- Knoldus Software LLC
- Localytics
- Magine TV
- MediaCrossing
- MyFitnessPal
- NASA JPL - Deep Space Network
- Netease
- NFLabs
- Nokia Solutions and Networks
- NTT DATA
- Nube Technologies
- Ooyala, Inc.
- Opentable
- PanTera
- Peerialism
- PlanBMedia
- PredicitionIo
- Premise
- Quantifind
- Radius Intelligence
- Real Impact Analytics
- RocketFuel
- RONDHUIT
- Sailthru
- Samsung Research America
- Shopify
- Simba Technologies
- Sinnia
- SK Telecom
- Socialmetrix
- Sohu
- Stanford DAWN
- Stratio
- Taboola
- Techbase
- Tencent
- Tetra Concepts
- TrendMicro
- TripAdvisor
- truedash
- TruEffect Inc
- UC Santa Cruz
- University of Missouri Data Analytics and Discover Lab
- VideoAmp
- Vistar Media
- Yahoo!
- Yandex
- Zaloni

EMPRESAS UTILIZANDO O SPARK

amazon

AUTODESK.

ebay

NASA JPL

Nokia Siemens Networks

SAMSUNG RESEARCH AMERICA

tripadvisor®

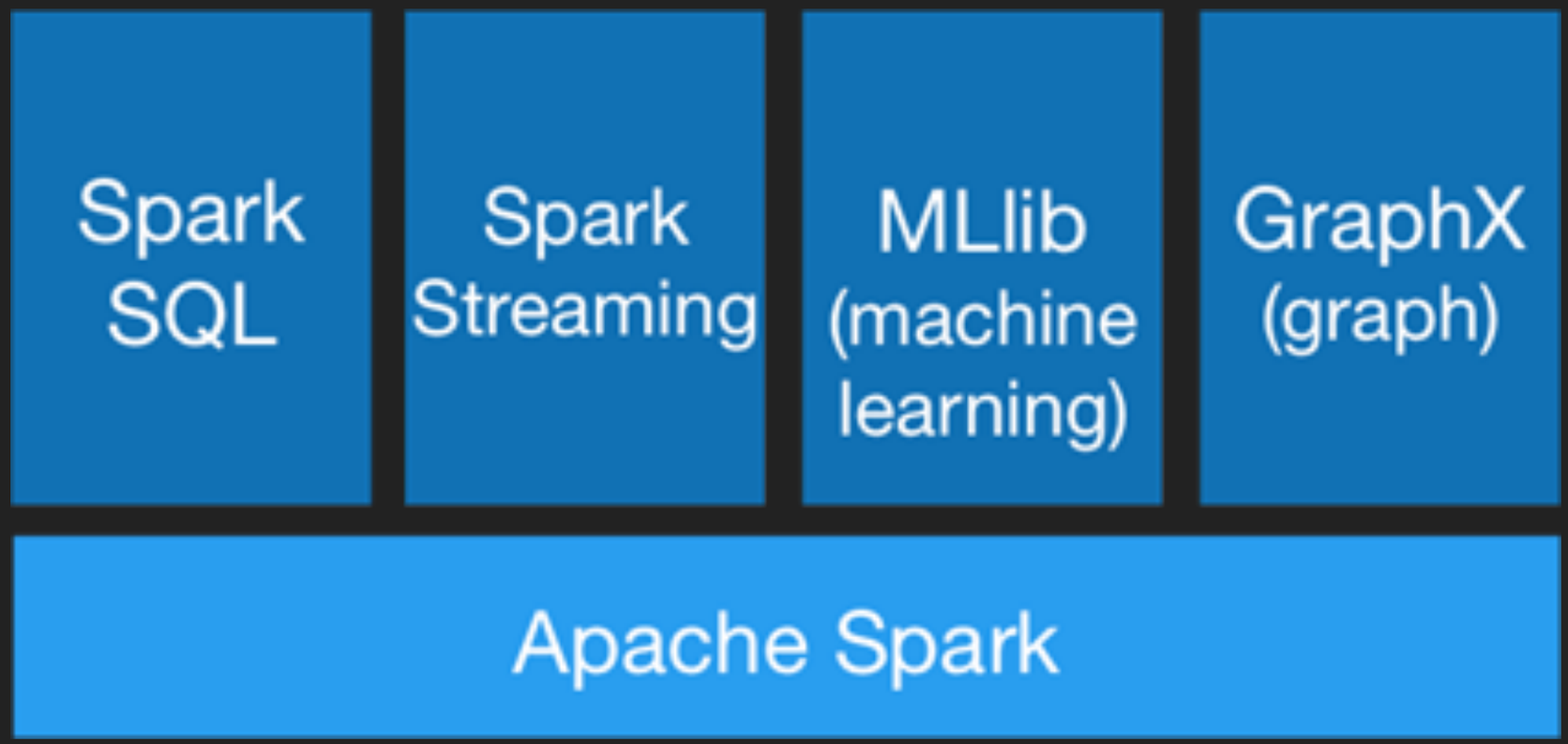
YAHOO!

CURIOSIDADES

- A maioria das empresas rodam com milhares de máquinas;
- A maior conhecida possui 8000;
- Se mostra trabalhar bem na casa dos PetaBytes;
- Ja foi usado pra ordenar 100TB, 3X mais rápido que o MapReduce do Hadoop;
- Ganhou o Daytona GraySort Benchmark de 2014 ordenando 1PB.

COMO ELE FUNCIONA?

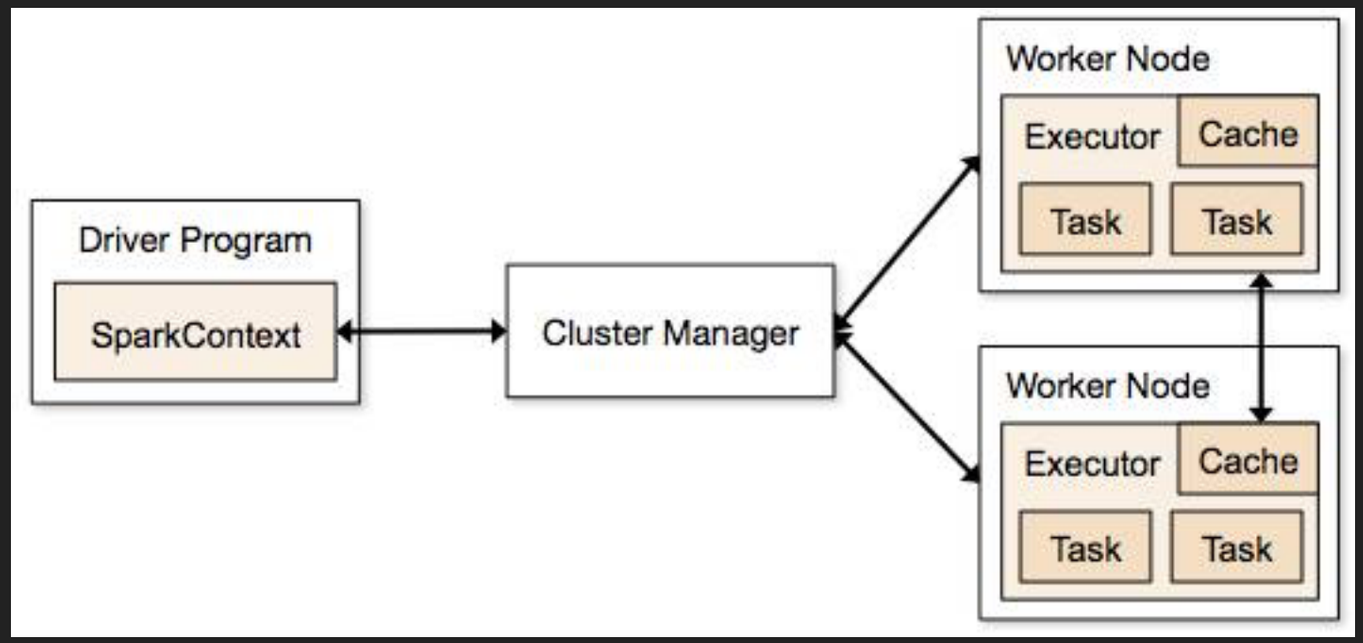
COMPONENTES



COMPONENTES

- O Spark Streamming, que possibilita o processamento de fluxos em tempo real;
- O GraphX, que realiza o processamento sobre grafos;
- O SparkSQL para a utilização de SQL na realização de consultas e processamento sobre os dados no Spark;
- A MLlib, que é a biblioteca de aprendizado de máquina, com deferentes algoritmos para as mais diversas atividades, como clustering.

ARQUITETURA



ARQUITETURA

- O Driver Program, que é a aplicação principal que realiza a criação das tarefas e é quem às envia para os Workers;
- O Cluster Manager é responsável por administrar as máquinas que serão utilizadas como workers;
- Os Workers, que são as máquinas que realmente executarão as tarefas que são enviadas pelo Driver Program. Se o Spark for executado de forma local, a máquina desempenhará tanto o papel de Driver Program como de Worker.

E O DEPLOY?

STANDALONE DEPLOY MODE

- https://spark.apache.org/docs/latest/spark-standalone.html

APACHE MESOS

- https://spark.apache.org/docs/latest/running-on-mesos.html

HADOOP YARN

- https://spark.apache.org/docs/latest/running-on-yarn.html

KUBERNETES

- https://spark.apache.org/docs/latest/running-on-kubernetes.html

AMAZON EC2

- https://github.com/amplab/spark-ec2#readme

APLICAÇÃO

ORDENAÇÃO

Maquinas	2	4	8
8gb	19 min	11 min	18 min KILLED
16gb	45 min	25 min	8 min
32gb	1h20min	37 min	17 min

ORDENAÇÃO