IASA Open Coffee: IoT & Big Data ...

Monday, June 29, 2015

IASA

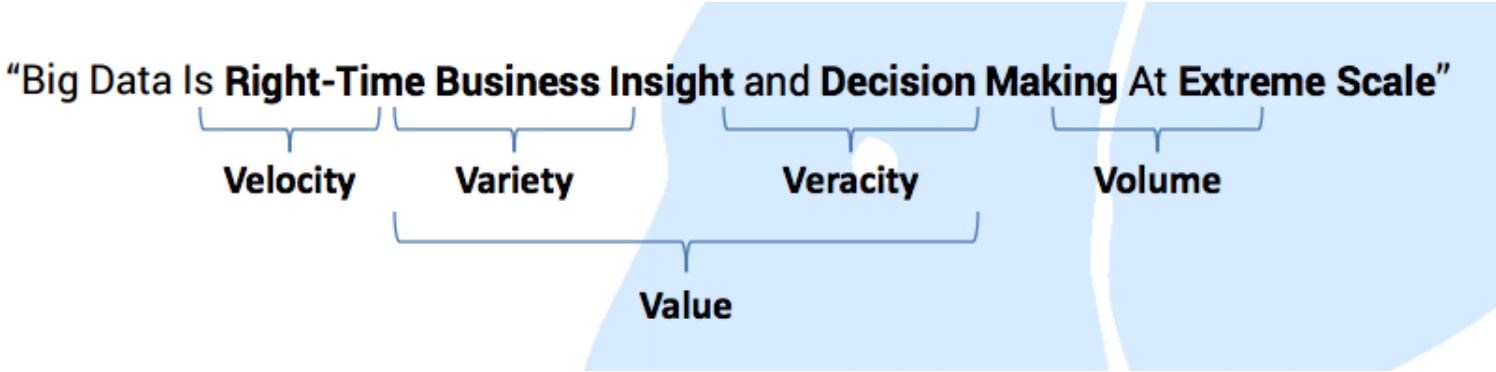
Agenda

- What is Big Data? Data warehousing ...
- Why? Business Benefits ...
- Specific business use cases: IoT, prediction ...
- Tools & Techniques (Hadoop, Spark, R, Python ...), Data Science ...
- Software & Infrastructure challenges
- Mini-Workshop using Spark
- See "Introduction to Apache Spark Workshop" @https://www.databricks.com/spark/developer-resources
- Additional resources, courses ...

What is Big Data? Data warehousing ...

Big data can be described by the following characteristics:

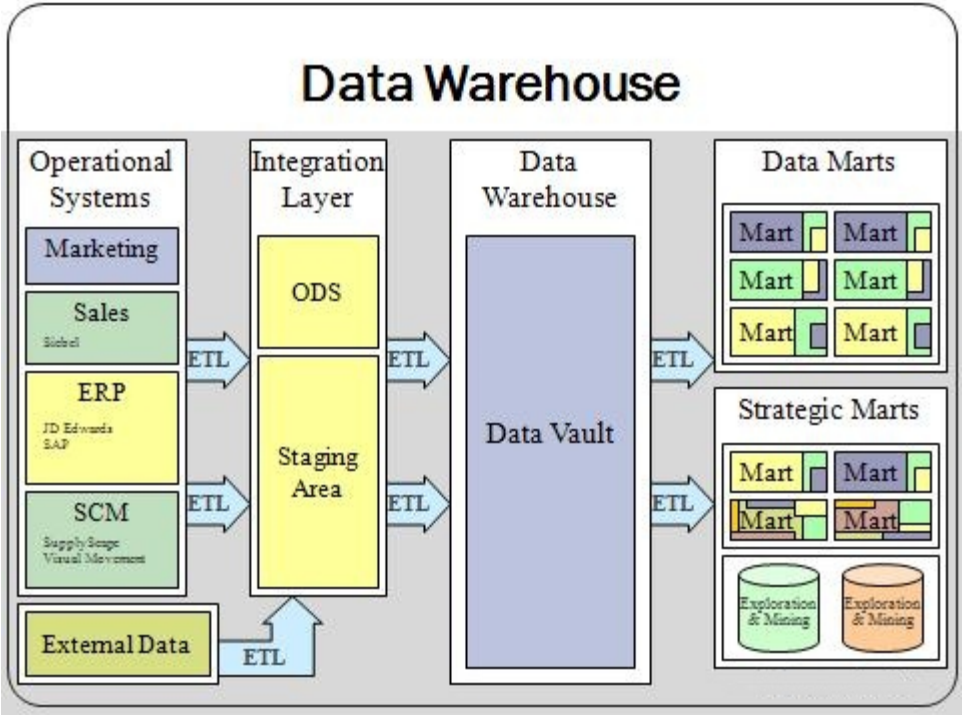
- Volume - from Terabytes to Petabytes
- Variety - data from many sources, cleaned and aggregated
- Velocity - can be generated real time from sensors, transactions, ...
- Variability - inconsistency in data, sensors down, data missing, ...
- Veracity - variable quality, missing data
- Complexity - data management, multiple sources, quality - all contribute to complexity



Data Warehousing

Old “Big Data”

- Clean(-er) sources of data
- Standard queries on data
- Monthly, weekly reports



Why? Business Benefits ...

Feedback from more data sources (streamed / live):

- Web clicks
- Log data
- Sensor updates

Uncover hidden behaviour

Gain competitive edge

Specific business use cases: IoT, prediction ...

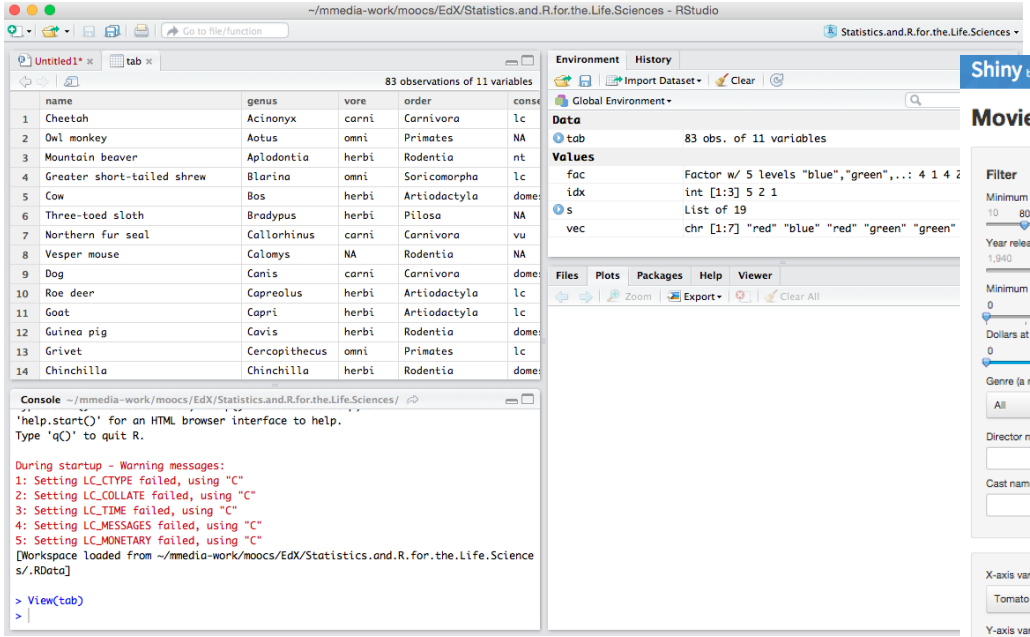
- Fraud detection
- Predicting Trends
- Real-time analytics
- Device measurements across millions of devices every 15 minutes (e.g. smart grid)
- User-generated content (e.g. Twitter/Facebook) Historical market transaction data (e.g. Netflix/Amazon)
- Transactions, click streams, and content for trend analysis
- Multiple real-time medical devices for patient assessment
- Filtering spam
- Normalizing data/Master Data Management
- Audit trails/data lineage
- Identifying new markets or demographics
- Netflix determining that the series House of Cards would succeed based on the viewing habits of their viewers
- Political campaigns using data to steer conversation with voters in near real-time

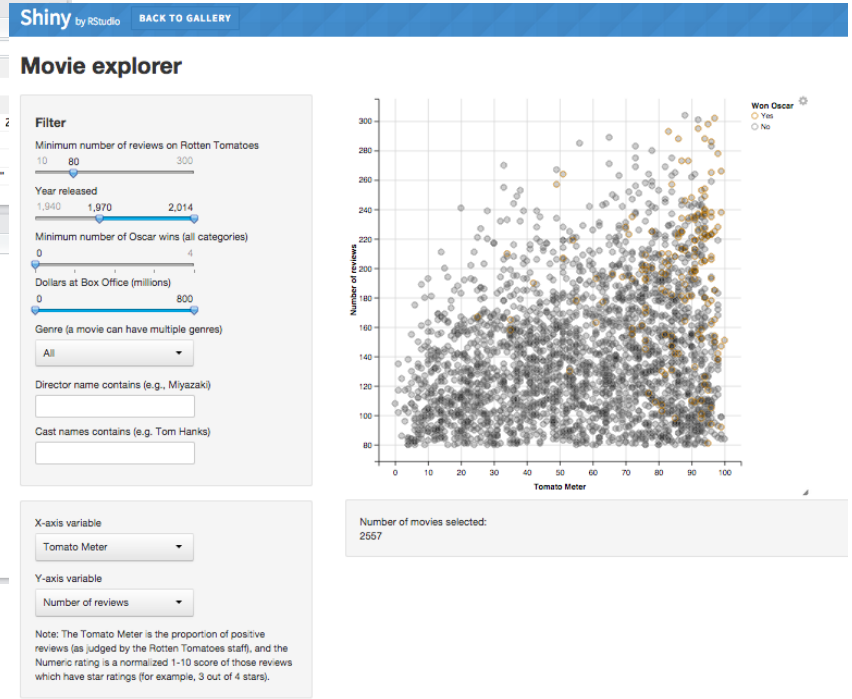
Tools and Techniques

Variety of tooling based on:

- R
- Python
- Scala / Spark
- Clojure / Incanter

R, RStudio, Shiny

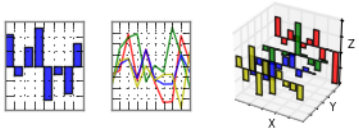




Python, Pandas ...

pandas

$y_{it} = \beta' x_{it} + \mu_i + \epsilon_{it}$



overview // get pandas // documentation // community // talks

Python Data Analysis Library

pandas is an open source, BSD-licensed library providing high-performance, easy-to-use data structures and data analysis tools for the Python programming language.

0.16.2 final (June 12, 2015)

This is a minor bug-fix release from 0.16.1 and includes a a large number of bug fixes along several new features, enhancements, and performance improvements. We recommend that all users upgrade to this version.

Highlights include:

- A new `pipe` method, see here
- Documentation on how to use numba with *pandas*, see here

See the full Whatsnew

For binaries and source archives of v0.16.2 final, see the GitHub Releases

VERSIONS

Release
0.16.2 - June 2015
download // docs // pdf

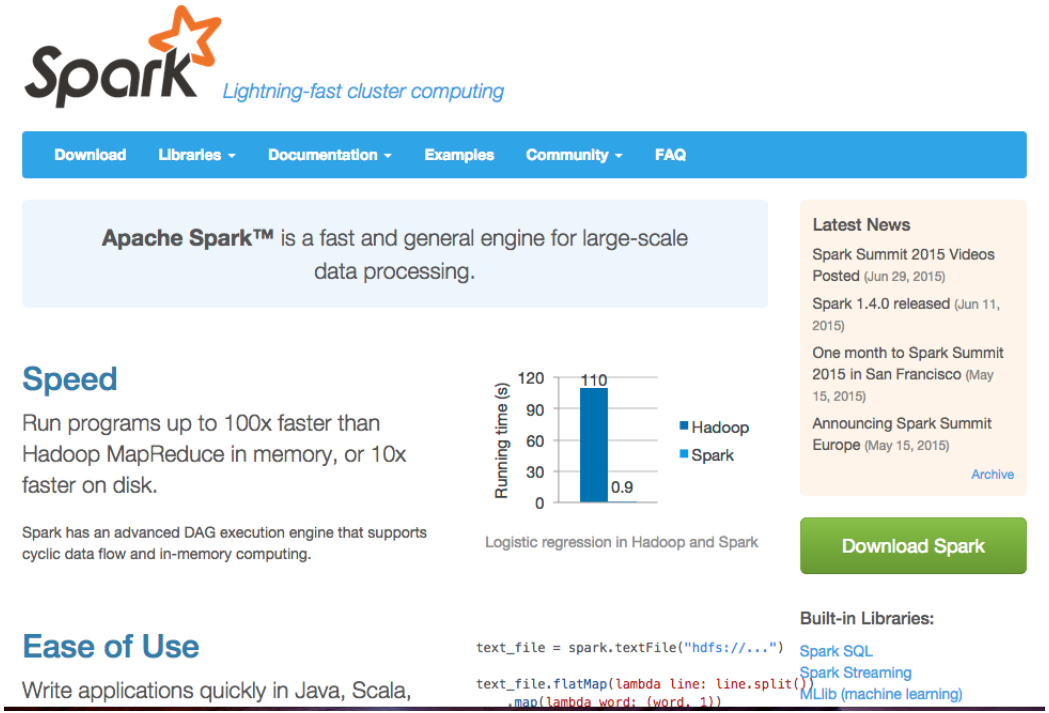
Development
0.17.0 - July 2015
github // docs

Previous Releases
0.16.1 - download // docs // pdf
0.16.0 - download // docs // pdf
0.15.2 - download // docs // pdf
0.15.1 - download // docs // pdf
0.15.0 - download // docs // pdf
0.14.1 - download // docs // pdf
0.14.0 - download // docs // pdf
0.13.1 - download // docs // pdf
0.13.0 - download // docs // pdf
0.12.0 - download // docs // pdf

Scala / Spark ...

Spark is now the Big Data tool of choice:

- Fast
- Clustering
- Parallel execution
- Easier programming mode
- Handles streaming data



Data Science ...

New discipline, wiki says:

“

Data Science is the extraction of knowledge from large volumes of data that are structured or unstructured,[1][2] which is a continuation of the field data mining and predictive analytics, also known as knowledge discovery and data mining (KDD). "Unstructured data" can include emails, videos, photos, social media, and other user-generated content. **Data Scientists** are qualified people with strength and patience to tunnel through mountains of information and the technical skills in writing algorithms to extract insights from these troves of information.

”

Software & Infrastructure challenges

… to be added later …

… clustering, compute on demand, cloud, data volumes, ...

Mini-Workshop using Spark

See "Introduction to Apache Spark Workshop" @ https://www.databricks.com/spark/developer-resources

Additional resources, courses ...

Check out EdX and Coursera:

Courses

This specialization covers the concepts and tools you'll need throughout the entire data science pipeline, from asking the right kinds of questions to making inferences and publishing results. The Specialization concludes ... Show more »

1. The Data Scientist's Toolbox
2. R Programming
3. Getting and Cleaning Data
4. Exploratory Data Analysis
5. Reproducible Research
6. Statistical Inference
7. Regression Models
8. Practical Machine Learning
9. Developing Data Products
10. Data Science Capstone

Course 1
The Data Scientist's Toolbox

Start Now - €26

Upcoming Session: June 22 201:

Duration: 4 weeks

Estimated Workload: 1-4 hours/week

Upon completion of this course you will be able to identify and classify data science problems. You will also have created your Github account, created your first repository, and pushed your first markdown file to your account.

edX HOW IT WORKS COURSES SCHOOLS & PARTNERS

Viewing 97 courses matching

data

"data"

VERIFIED
TeachersCollegeX
BDE1x
Big Data in Education
Starting Soon
Starts: July 1, 2015

VERIFIED
HarvardX
GSE3x
Introduction to Data Wise: A Collaborative Process to Improve Learning & Teaching
Current
Starts: Self-Paced

VERIFIED
DelftX
EX101x
Data Analysis: Take It to the MAX()
Upcoming
Starts: September 2015

VERIFIED
UC BerkeleyX
CS100.1x
Introduction to Big Data with Apache Spark

UTArlingtonX
LINK5.10x
Data, Analytics and Learning

VERIFIED
HarvardX
PH525.8x
Case study: DNA methylation data analysis