Data engineering on Yelp Datasets

Faculty: Michael Enudi

About Me.

- Lives and works in Johannesburg, South Africa
- Senior Software engineer with over 10 years of working experience writing enterprise java applications, architecting data solutions.
- Cloudera Certified Spark and Hadoop Developer.
- Oracle Certified SQL Expert
- Oracle Certified Java Master
- Sun Certified Java Business Component Dev.
- Sun Certified Java Programmer
- Big data enthusiast

LinkedIn → https://www.linkedin.com/in/michaelenudi

What is data engineering

Data engineering involves the science and art of building, automating and maintaining a data infrastructure for an organization. It involves many skills from data analysis, software development, data integration, warehousing, data governance, visualization, optimization and network architecture.

Worth reading: http://www.mastersindatascience.org/careers/data-engineer/

Data engineers

Designs, builds and manages the information or data software infrastructure and pipeline (data at rest and data in motion.

Data scientist

Applies data science and analytic results to critical business issues - helping an organization turn data into information - information into knowledge and insights - and valuable, actionable insights into better decision making and game changing strategies.

Data Engineering involves

- Data acquisition
- Storage
- Process Optimization
- Data pipeline automation
- Data governance
- Testing and Deployment
- Working with data scientist
- Understanding business value from data
- Software development
- Security and privacy

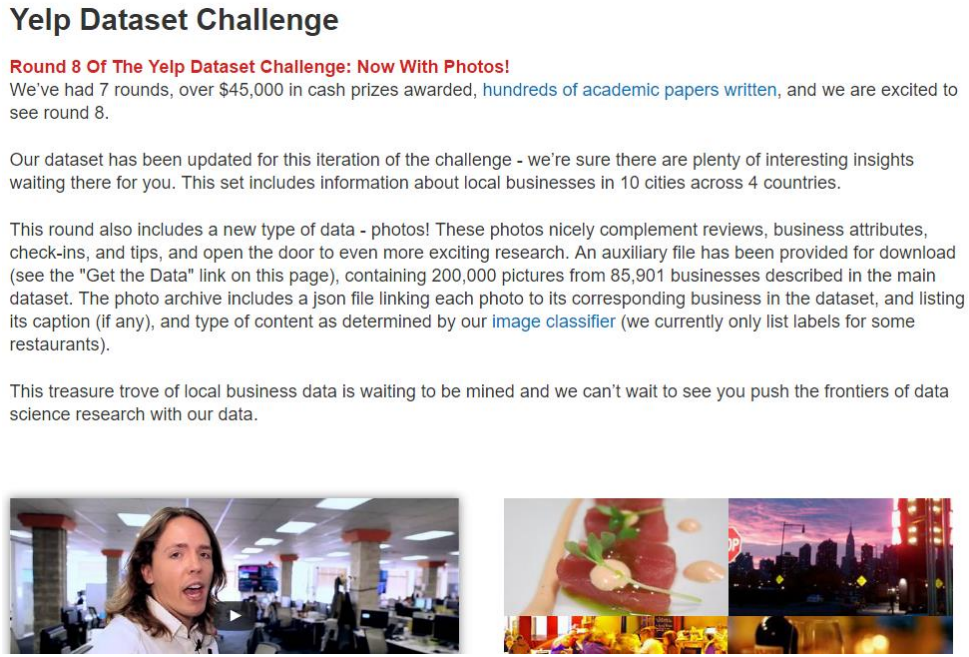
Yelp

Yelp is an American multinational corporation headquartered in San Francisco, California. It develops, hosts and markets Yelp.com and the Yelp mobile app, which publish crowd-sourced reviews about local businesses, as well as the online reservation service Yelp Reservations and online food-delivery service Eat24. The company also trains small businesses in how to respond to reviews, hosts social events for reviewers, and provides data about businesses, including health inspection scores. (Wikipedia: https://en.wikipedia.org/wiki/Yelp)

Yelp data challenge

Years ago, Yelp (NYSE: YELP) made the audacious move to prepare and make avialable dataset on businesses, reveiws, users, check-ins and tips available on the internet for the purpose of encourage data science innovation to interested audience in the field with over $45,000 in cash prizes awarded and over hundreds of academic papers written as a result.
The current round of the competition ends on the 31st December, 2016.

See more https://www.yelp.com/dataset_challenge



Yelp Dataset Challenge

Round 8 Of The Yelp Dataset Challenge: Now With Photos!
We've had 7 rounds, over $45,000 in cash prizes awarded, hundreds of academic papers written, and we are excited to see round 8.

Our dataset has been updated for this iteration of the challenge - we're sure there are plenty of interesting insights waiting there for you. This set includes information about local businesses in 10 cities across 4 countries.

This round also includes a new type of data - photos! These photos nicely complement reviews, business attributes, check-ins, and tips, and open the door to even more exciting research. An auxiliary file has been provided for download (see the "Get the Data" link on this page), containing 200,000 pictures from 85,901 businesses described in the main dataset. The photo archive includes a json file linking each photo to its corresponding business in the dataset, and listing its caption (if any), and type of content as determined by our image classifier (we currently only list labels for some restaurants).

This treasure trove of local business data is waiting to be mined and we can't wait to see you push the frontiers of data science research with our data.

Yelp dataset & schema

Interested areas of research

- Cultural Trends
- Location Mining and Urban Planning
- Seasonal Trends
- Infer Categories
- Natural Language Processing (NLP)
- Changepoints and Events
- Social Graph Mining

Domain

- Business
- Review
- User
- check-in
- tip
- photo

https://www.yelp.com/dataset_challenge

Data Ingestion?

Our Yelp dataset has already been provided and to an extent prepared.

So we are not focusing on data ingestion in this meeting

Decision Making: Data format and storage

Data format

- Text vs Binary

Container file format

- Row vs Columnar access

Compression ?

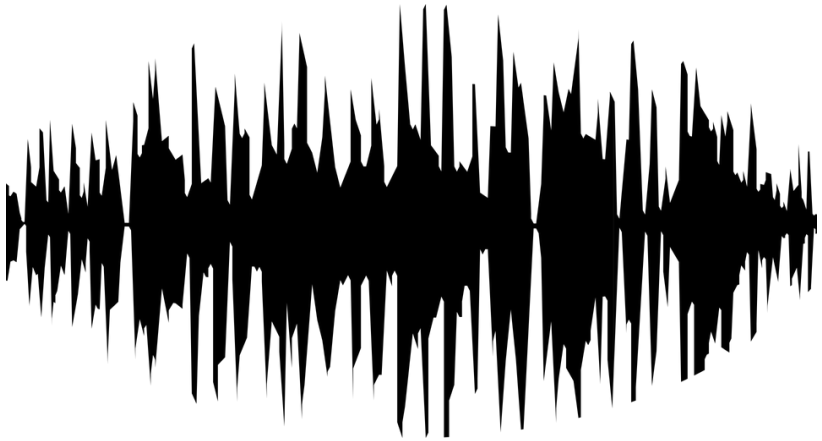
- Codec

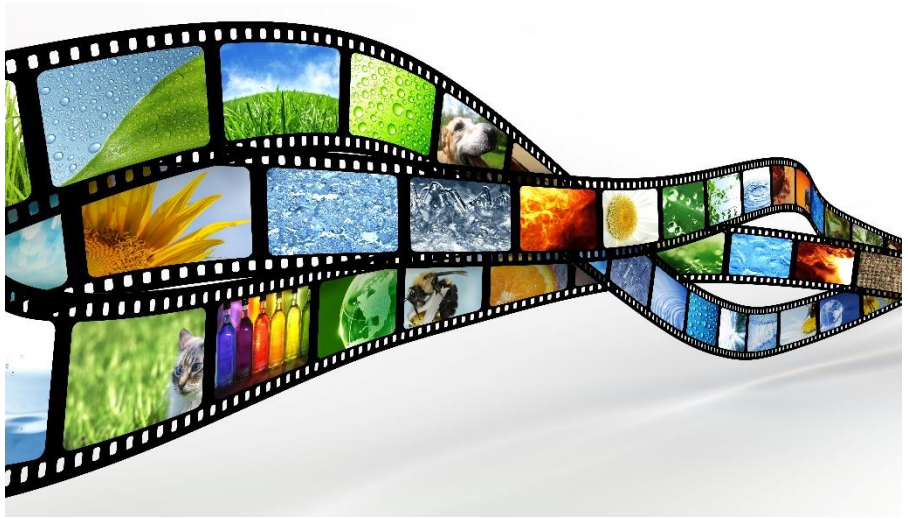
Processing Tools

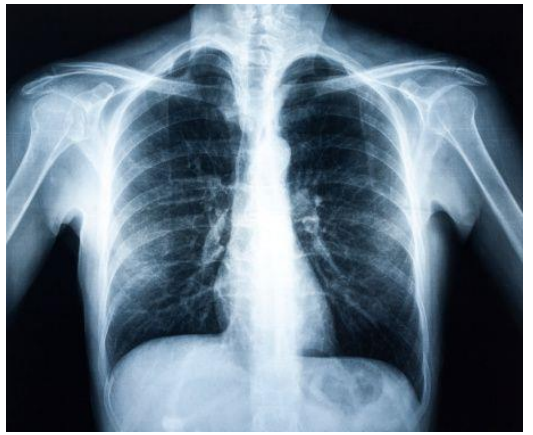
Decision Making: Data processing tools

- Do we need further processing to get value of out the data or can we somehow use the data as is?
- What is the nature of teams across the organization and SLAs guide the use of tools?
- What is the data governance procedures in the organization?

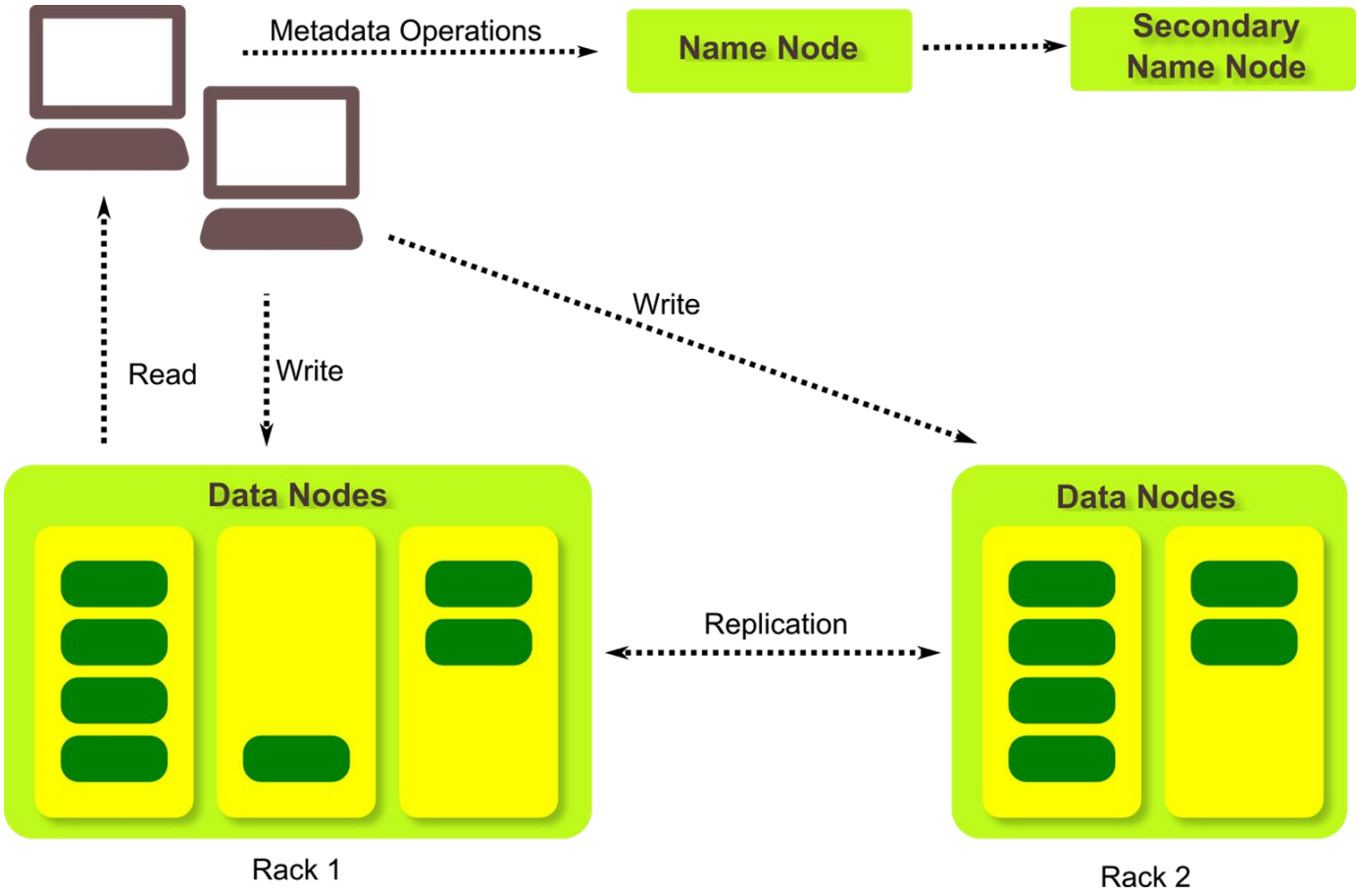
Decision Making: Processing binary content



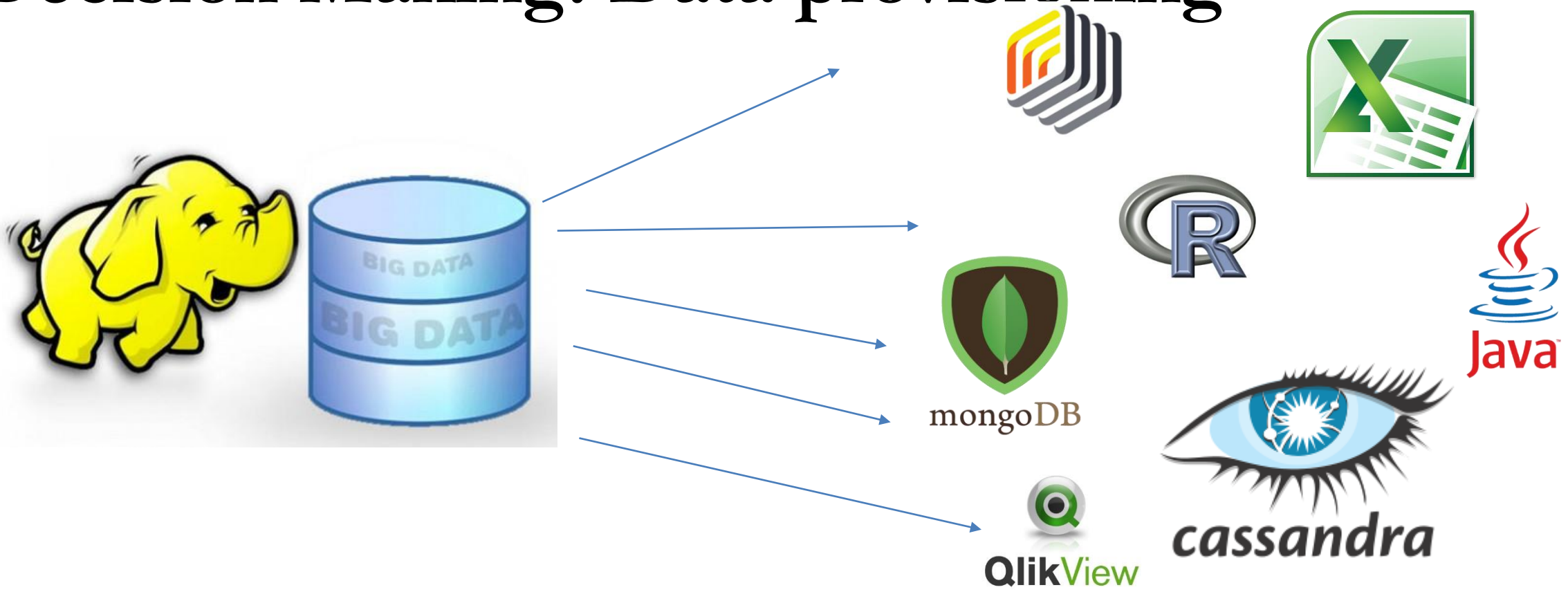




Decision Making:
Hadoop small file problem



Decision Making: Data provisioning



- Hive/Impala or other MPP database systems
- Sqoop to RDBMS
- Sqoop to NoSQL (Cassandra, Hbase, MongoDB, etc)
- Serve using application API (REST or Soap web services)

Decision Making:
Other alternative platforms

Are there any better way to do this? Online vs Batch

Think in Graphs – Neo4J

Implementation

Requirements

- Cloudera Quickstart VM 5.7 or 5.8
- Scala SDK and Runtime
- Scala build tool

Thank You