End to End Automated Data Pipeline of Yelp Reviews

Pranav Prajapati
Instructor: Joseph Morabito

STEVENS INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®
Business Intelligence & Analytics

Introduction

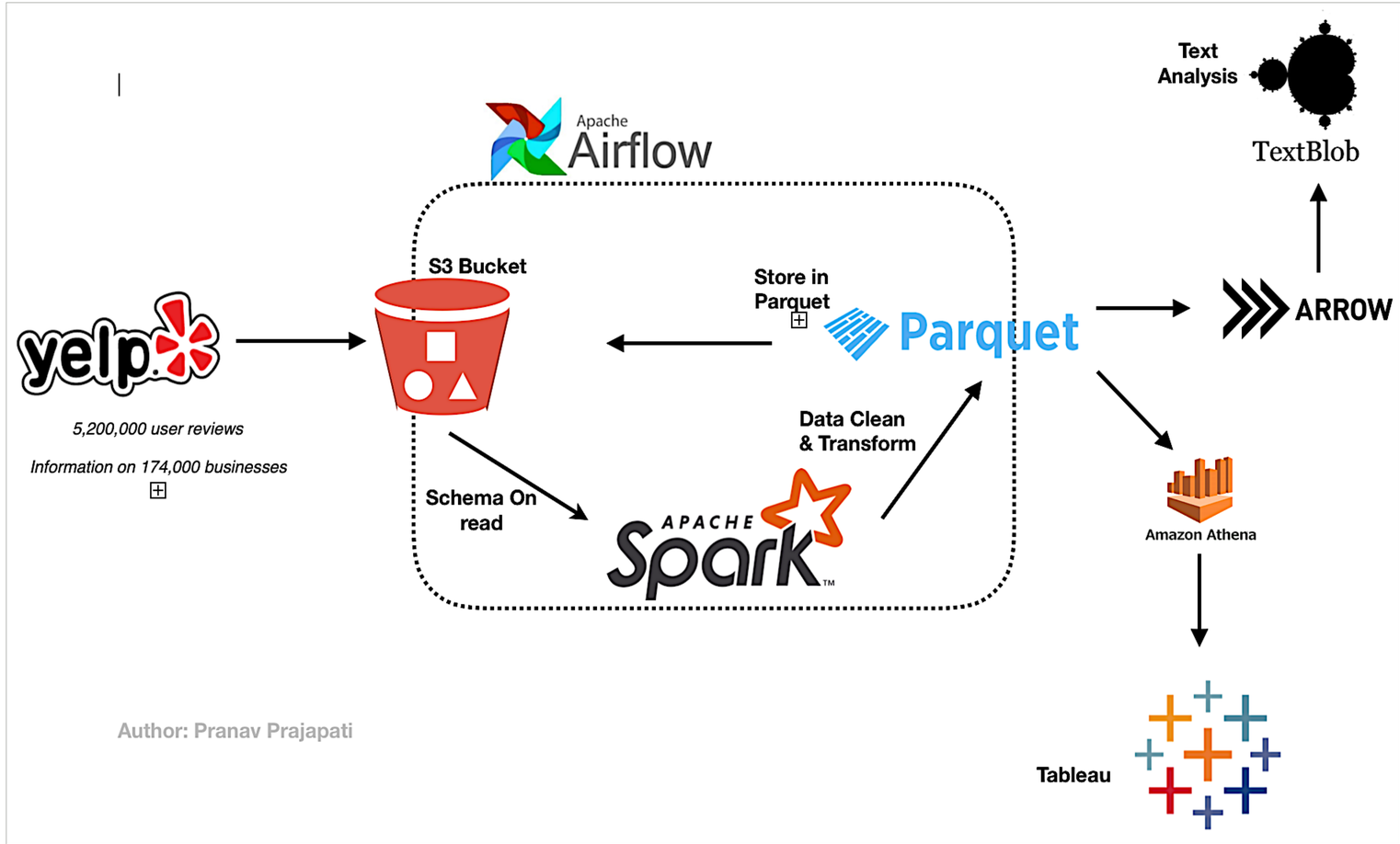
- The Star treatment – Chefs go all out to impress reviewers
- 137 million active Yelp Reviews

Evolution of Data Warehouse and why Data Lakes ?

- Abundance of unstructured data
- Rise of Big Data technologies
- Advanced Analytic Capabilities
- Schema on Read

The data

- Deep JSON and CSV file formats
- Business, Review, Tip, Check-in, User data
- Restaurants, Compliment tables created

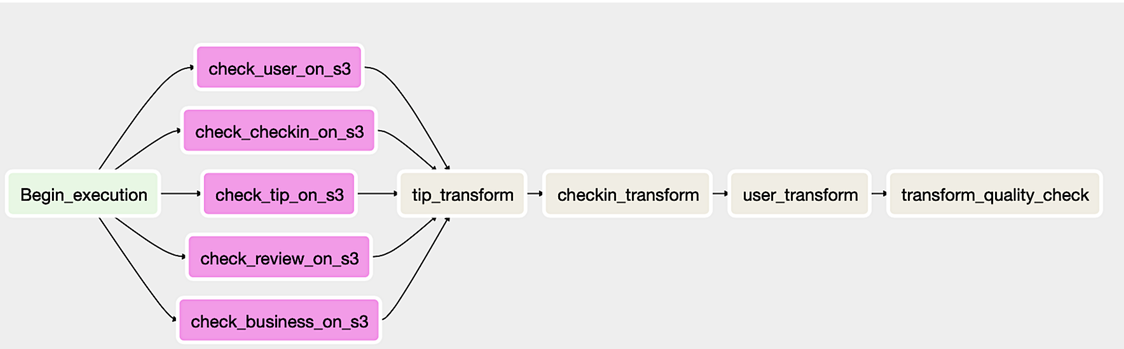


Overplanned Analytics Initiatives Are Doomed to Fail

Airflow and DAGs

- Use airflow to author workflows as directed acyclic graphs (DAGs) of tasks
- The rich user interface makes it easy to visualize pipelines running in production, monitor progress, and troubleshoot issues when needed
- We can automate tasks and schedule them whenever we want – Endless Possibilities !
- Essential Data Quality checks are added

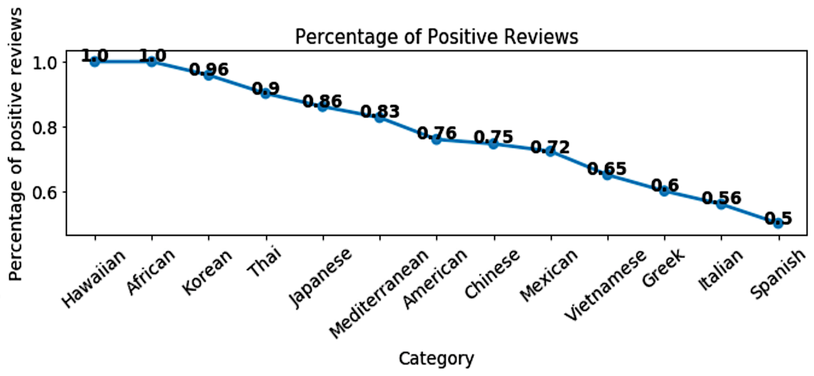
First DAG

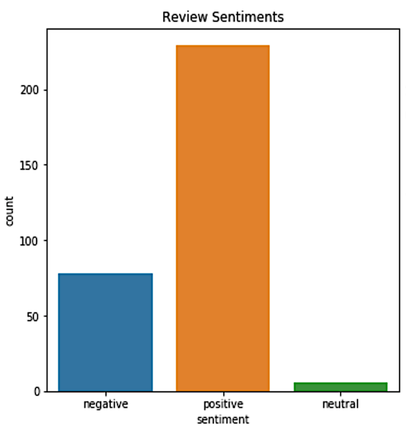


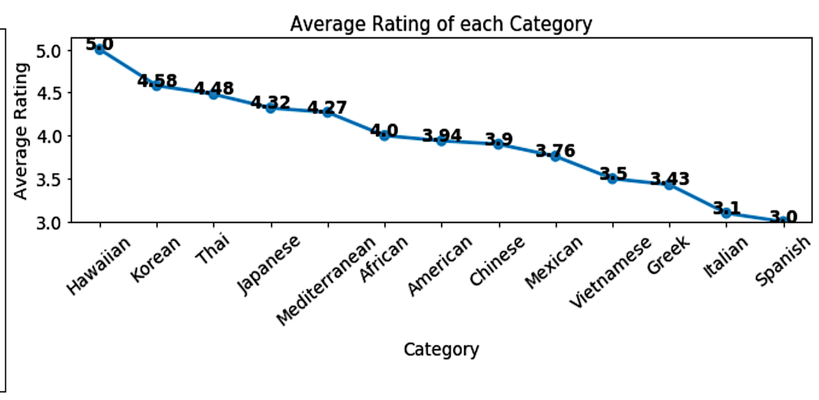
Second DAG

Analysis of the Restaurant Reviews

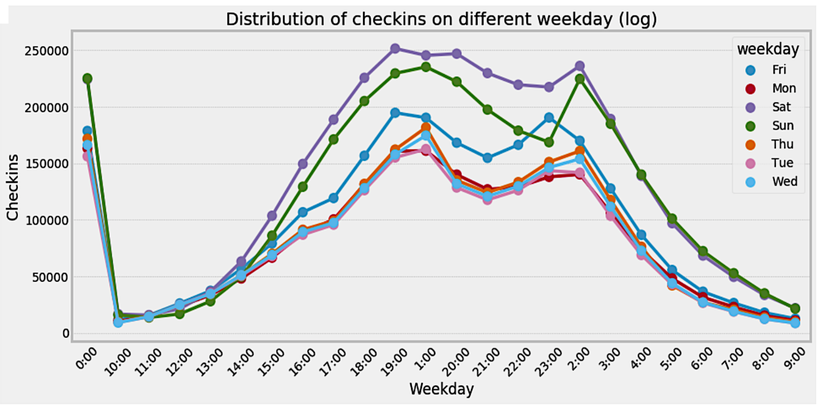
- Do people always Rate positively ?
- Which Cuisines are liked the most?
- Sentiment towards specific cuisines







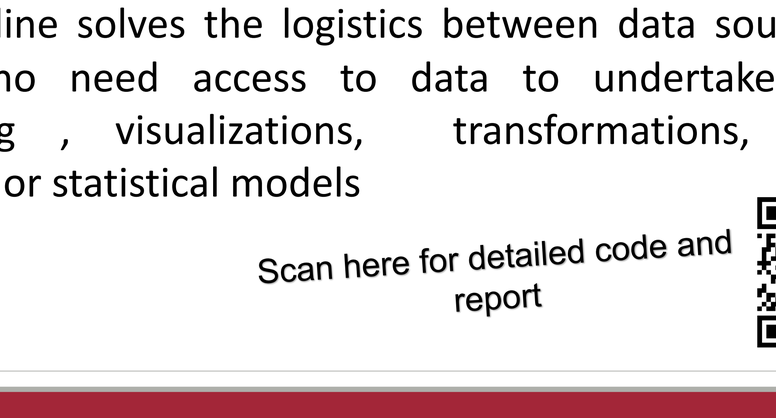
- Daily check-ins of users
- Partitioning the Data on S3 gives performance boost



59%** is the amount by which revenue of restaurants have increased when the average ratings increase by **1 star

Connect to your Dashboard

- Configure Amazon Athena Connector with Tableau to read the Data
- Create Business Intelligence Reports
- Present information in a more interactive manner



Conclusion

This Pipeline solves the logistics between data sources and those who need access to data to undertake further processing , visualizations, transformations, routing, reporting or statistical models

Scan here for detailed code and report