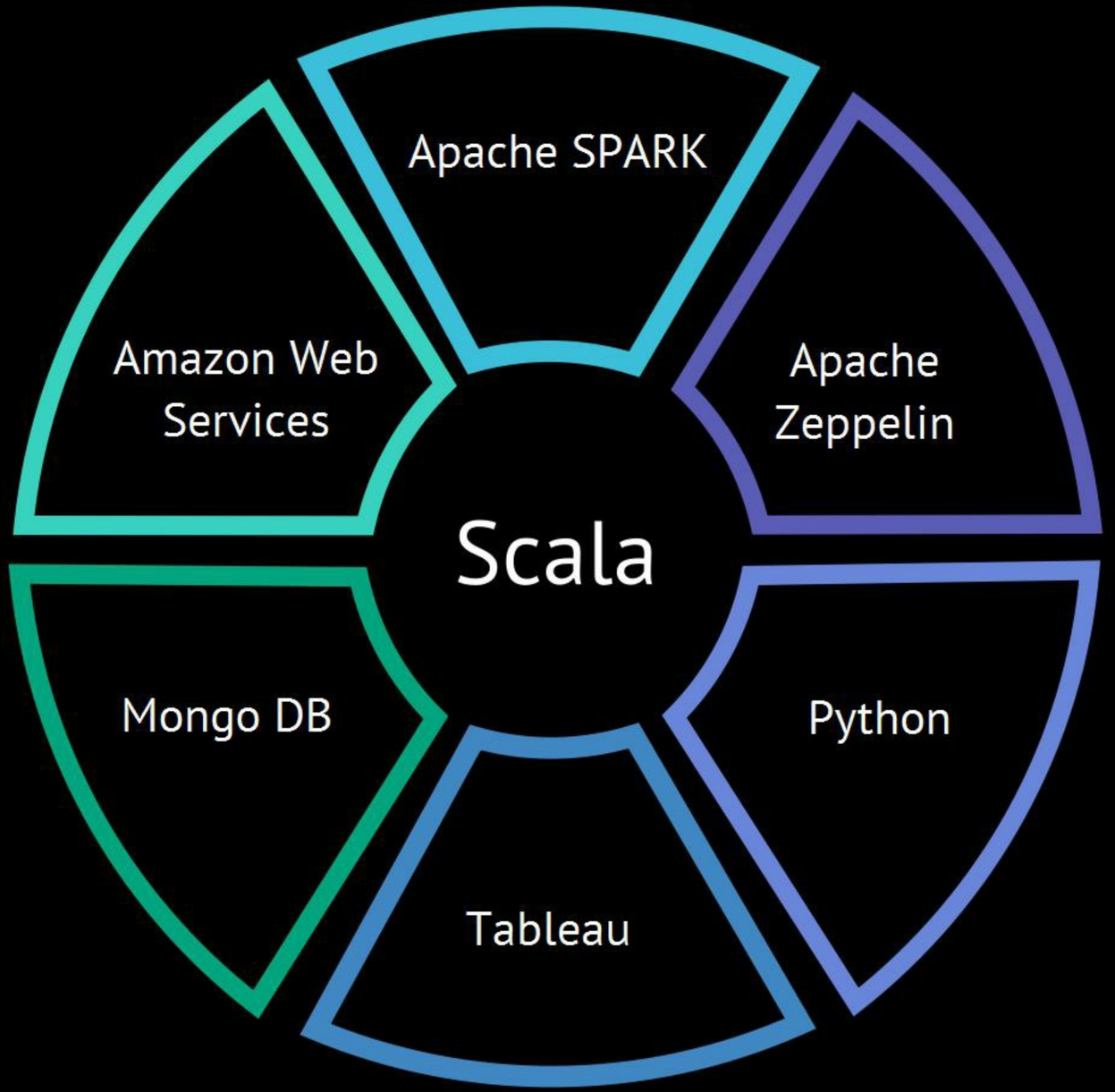
CRITIQUE MINING (E-COMMERCE REVIEWS)

Bala Gopalakrishnan

Malika Thapa

Rachan Hegde



AGENDA

01 Data Set

02 Goals

03 Scala Components

04 Big Data Analytics

05 Future Scope

DATA SET

Amazon Product Reviews

- Information about the products and reviews are in JSON format
- The volume of data set under consideration is ~10GB (> 10 Million Records)
- The details are grouped into the following data sets
  - Product Metadata (asin, product title, brand ...)
  - Product Reviews (asin, review, rating ...)

GOALS

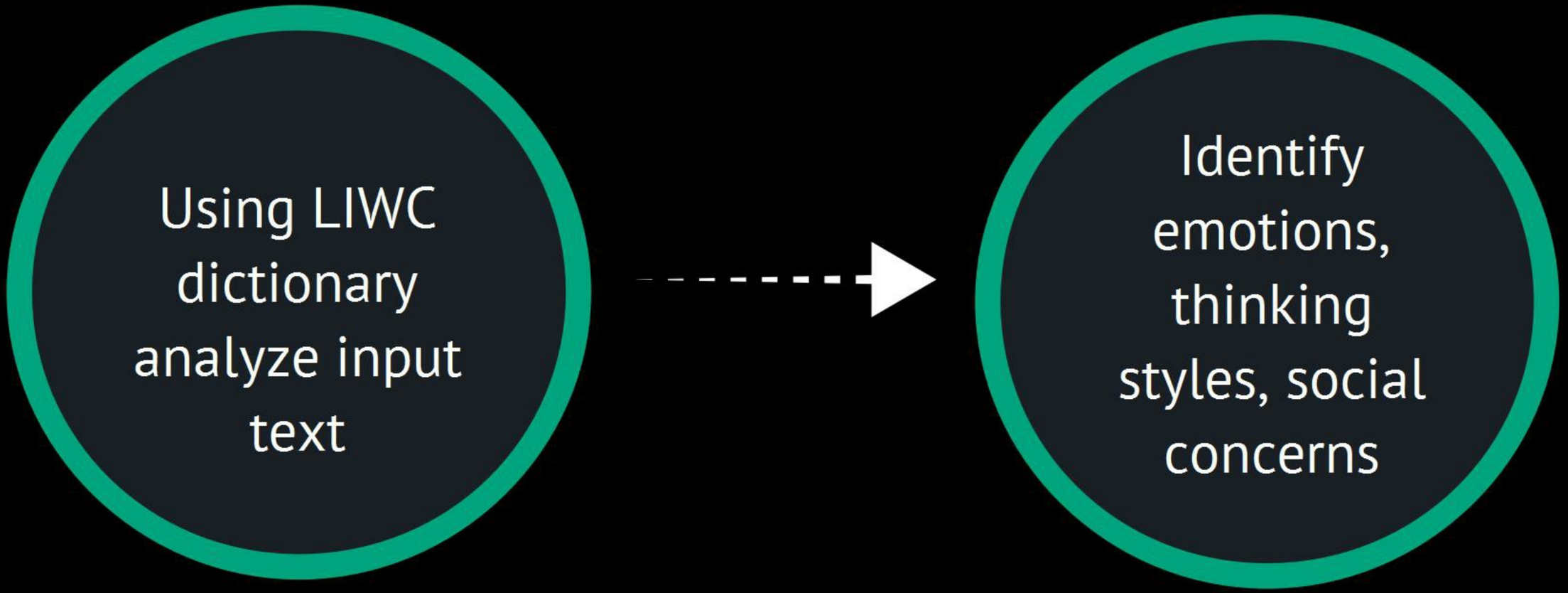
IT

- Use NoSQL to store product reviews (JSON)
- Create a LIWC dictinonay
- Compare reviews with LIWC dictionary to generate features set for ML
- Predict the actual rating using Spark ML
- Parallelize using SPARK

Business

- Find products with better customer satisfaction
- Neglect Fake Reviews
- Dashboards for informed decisions

LIWC (LINGUISTIC INQUIRY AND WORD COUNT)



SCALA COMPONENTS

01

NoSQL Data Loader

Configurable data loader (Import, Update, Upsert) for MongoDB

02

Search Index Generator

Converts LIWC text dictionary set into TRIE to improve search performace

03

LIWC Feature Generator

Enterprise will continue to exist in a foreseeable future.

04

ML Pipelines

Expenditure which brings into existence asset or benefit of a long term nature.

USE CASES

Actor: Data Engineer/ Scientist

- Load data into MongoDB (NoSQL Data Loader)
- Convert LIWC dictionary into TRIE (Search Index Generator)
- Create Feature Set for ML (LIWC Feature Generator)
- Apply ML on feature set to predict actual rating (ML Pipeline/ Zeppelin)

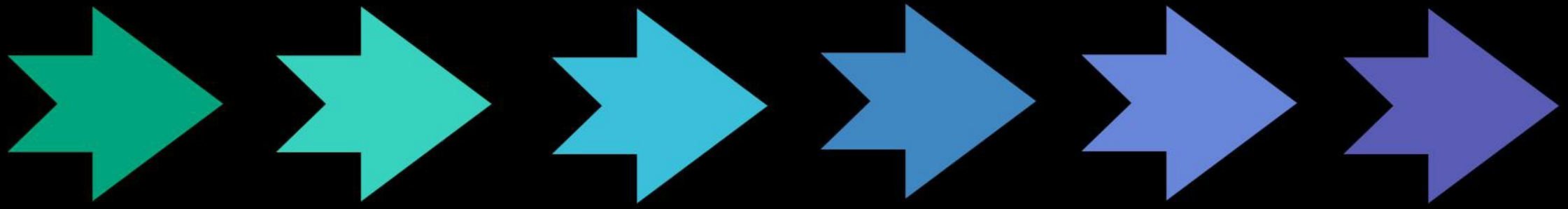
Actor: Business Executive

- Analyze brand performance based on average review rating and categories (Tableau)

Actor: Inventory Manager

BIG DATA ANALYTICS

Big Data Analytics



Load Data into Mongo DB
Scala
Mongo DB
Casbah
EmbedMongo

Analyze data for inconsistadncies
Mongo Shell
Mongo Compass
Tableau

Create LIWC Search Index
Scala
JSON Util

Using LIWC search index, create feature set for each review
Scala
JSON Util

Apply spark ML using feature set generated
Scala
SPARK ML
Regression
PCA Pipeline
Cross Validator
Metrics
Zeppelin

Update Mongo DB with features and predicted rating
Scala
Casbah
EmbedMongo

Future Scope

Aggregated Product Sentiments

Across
Dimensions

Product Recommender

Associative
Rule Mining

Seamless Integration between Modules

Integration is
currently manual

ACCEPTANCE CRITERIA

- Data loader module to parse JSON data and to import, update, upsert data into MongoDB
- Text analyzer module and feature set creation for ML (Map review text into onbr of the LIWC dimensions)
- Enable business users with dashboard to search for better performing brands in select categories
- Spark to parallelize the process and Spark ML to predict the star rating of reviews
- Setting a distributed environment using Amazon EMR to execute these process

Thank You !!!