SPARK+AI SUMMIT 2019

# Leveraging NLP and Deep Learning for Document Recommendation in the Cloud

Guoqiong Song, Intel

#UnifiedAnalytics #SparkAISummit

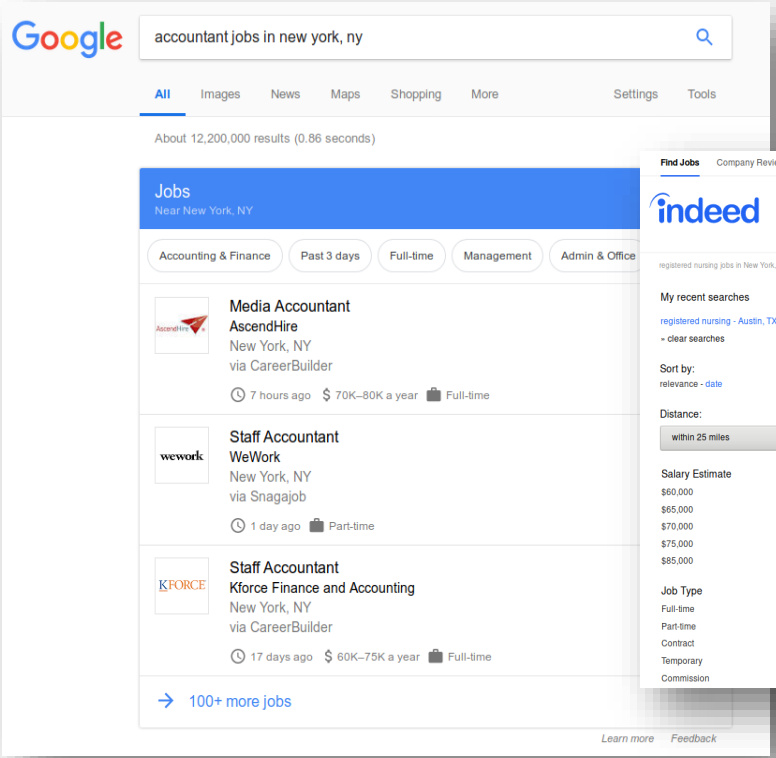
# Agenda

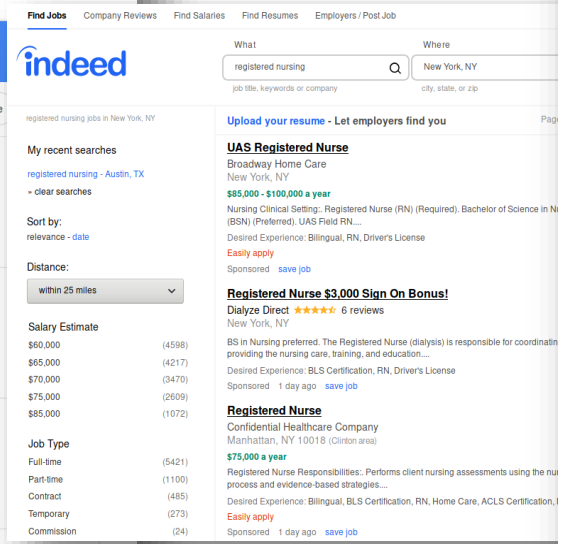
- Job/Resume Search Challenges and Opportunity
- Analytics Zoo and BigDL Overview
- Resume Search Analytics Zoo Solution
- Takeaways

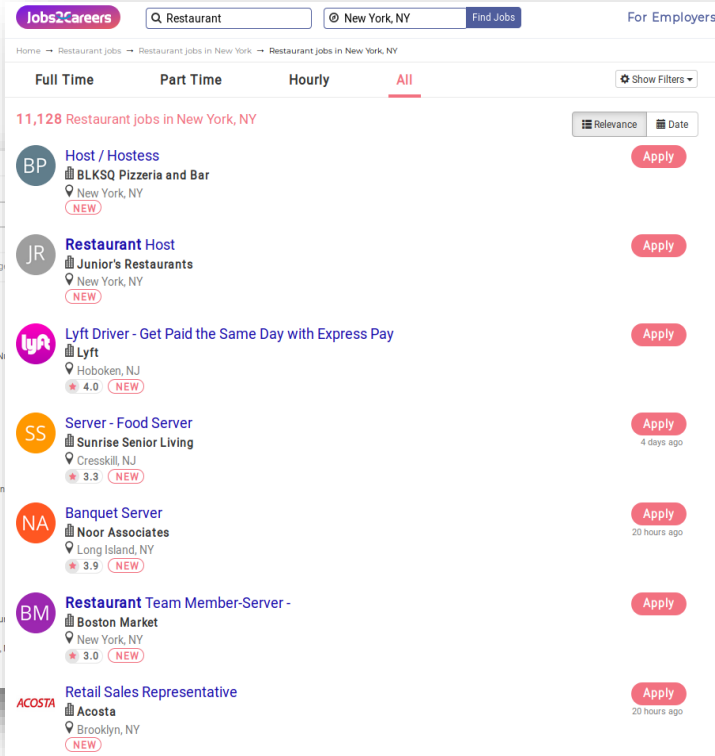
# Agenda

- **Job/Resume Search Challenges and Opportunity**
- **Analytics Zoo and BigDL Overview**
- **Resume Search Analytics Zoo Solution**
- **Takeaways**

# Job search


Google
accountant jobs in new york, ny
All
Images
News
Maps
Shopping
More
Settings
Tools
About 12,200,000 results (0.86 seconds)
Jobs
Near New York, NY
Accounting & Finance
Past 3 days
Full-time
Management
Admin & Office
Media Accountant
AscendHire
New York, NY
via CareerBuilder
7 hours ago
70K–80K a year
Full-time
Staff Accountant
WeWork
New York, NY
via Snagajob
1 day ago
Part-time
Staff Accountant
Kforce Finance and Accounting
New York, NY
via CareerBuilder
17 days ago
60K–75K a year
Full-time
100+ more jobs
Learn more
Feedback


Find Jobs
Company Reviews
Find Salaries
Find Resumes
Employers / Post Job
indeed
What
registered nursing
Where
New York, NY
My recent searches
registered nursing - Austin, TX
clear searches
Sort by:
relevance - date
Distance:
within 25 miles
Salary Estimate
$60,000 (4598)
$65,000 (4217)
$70,000 (3470)
$75,000 (2609)
$85,000 (1072)
Job Type
Full-time (5421)
Part-time (1100)
Contract (485)
Temporary (273)
Commission (24)
Upload your resume - Let employers find you
UAS Registered Nurse
Broadway Home Care
New York, NY
$85,000 - $100,000 a year
Easily apply
Registered Nurse $3,000 Sign On Bonus!
Dialyze Direct
6 reviews
New York, NY
Registered Nurse
Confidential Healthcare Company
Manhattan, NY 10018
$75,000 a year
Easily apply


Restaurant
New York, NY
Find Jobs
For Employers
Full Time
Part Time
Hourly
All
Show Filters
11,128 Restaurant jobs in New York, NY
Relevance
Date
Host / Hostess
BLKSQ Pizzeria and Bar
New York, NY
Restaurant Host
Junior's Restaurants
New York, NY
Lyft Driver - Get Paid the Same Day with Express Pay
Lyft
Hoboken, NJ
Server - Food Server
Sunrise Senior Living
Cresskill, NJ
Banquet Server
Noor Associates
Long Island, NY
Restaurant Team Member-Server -
Boston Market
New York, NY
Retail Sales Representative
Acosta
Brooklyn, NY
Apply

# Personalize Results Value

## Job Seekers

Find the right job faster

Resume

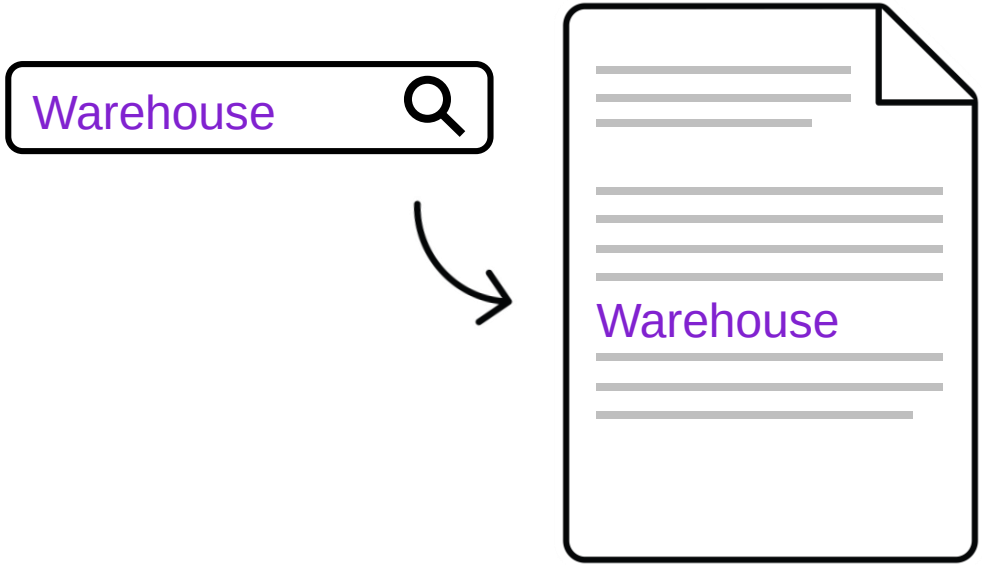
## Employers

Find the right person

Job description

# Traditional Information Retrieval Sufferings

Solution challenges: stemming, synonyms, ontologies, sensitivity



# Stemming

Accountant =!= Accounting

# Stemming Solution

Accountant =!= Accounting

Account~~ant~~ = Account~~ing~~

# Stemming Sufferings

Accountant =!= Accounting

---

Accountant = Accounting

---

Accountant = Accounting = Account Representative

# Synonyms

Registered Nurse =!= RN

# Synonyms Solution

| | | |
|---|---|---|
| Registered Nurse | =!= | RN |
| Registered Nurse | = | RN → registered nurse |

# Synonym Sufferings

| | | |
|---|---|---|
| Registered Nurse | =!= | RN |
| Registered Nurse | = | RN → registered nurse |

. . .

# Ontologies

Dishwasher =!= Back of House

# Ontologies Solution

Dishwasher =!= Back of House

Dishwasher Back of House

Restaurant = Restaurant

# Ontologies Sufferings

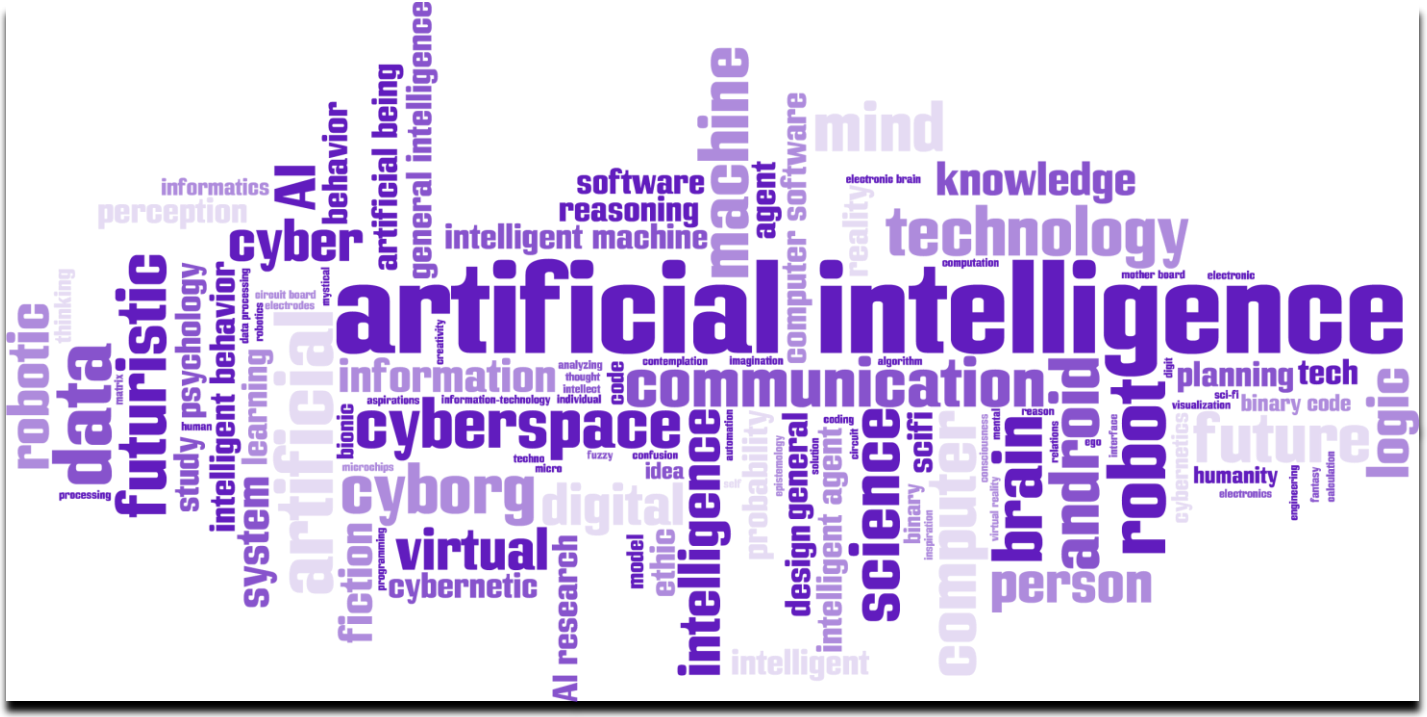
Dishwasher =!= Back of House

Dishwasher Back of House

Restaurant = Restaurant

. . .

# Specificity Suffering



# Personalize Results Value

## Job Seekers

Find the right job faster

Resume

## Employers

Find the right person

Job description

# Agenda

- Job/Resume Search Challenges and Opportunity
- **Analytics Zoo and BigDL Overview**
- Resume Search Analytics Zoo Solution
- Takeaways

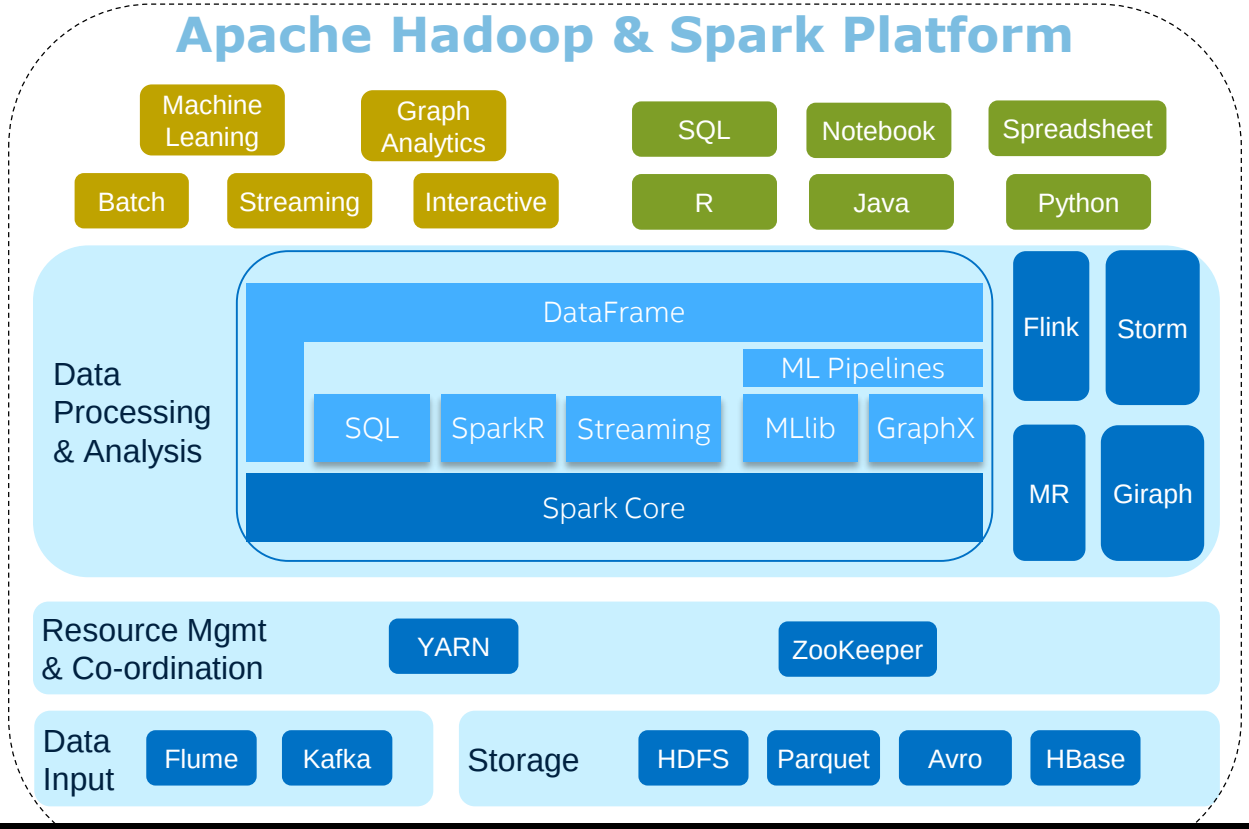
# AI on Spark


Distributed, High-Performance
Deep Learning Framework
for Apache Spark
https://github.com/intel-analytics/bigdl

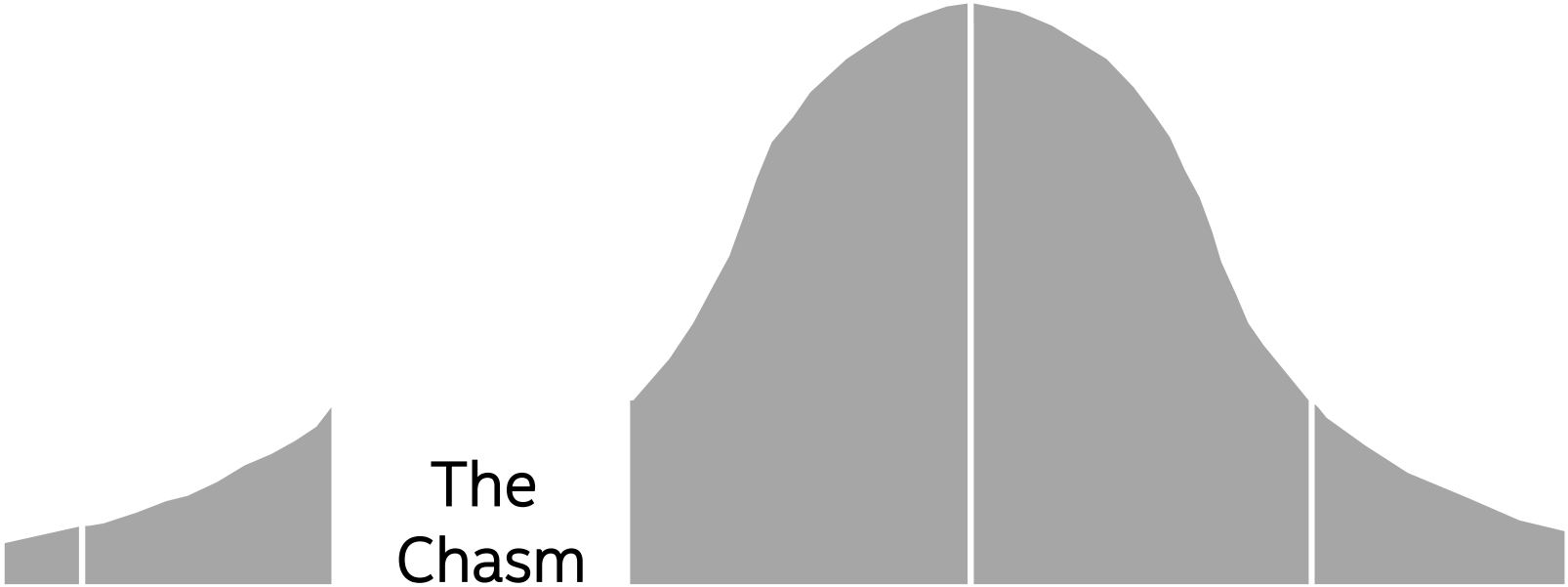

Distributed TensoRflow, Keras and BigDL on Spark
Reference Use Cases, AI Models,
High-level APIs, Feature Engineering, etc.
https://github.com/intel-analytics/analytics-zoo

**Unifying Analytics + AI on Apache Spark**

# Unified Big Data Analytics Platform



# Chasm b/w Deep Learning and Big Data Communities




Deep learning experts

Average users (big data users, data scientists, analysts, etc.)

# Bridging the Chasm

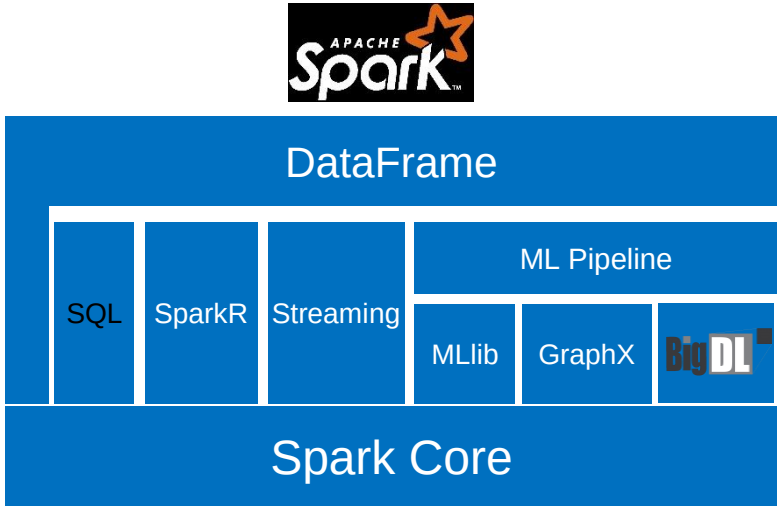
## Make deep learning more accessible to big data and data science communities

- **Continue the use of familiar SW tools and HW infrastructure to build deep learning applications**
- **Analyze "big data" using deep learning on the same Hadoop/Spark cluster where the data are stored**
- **Add deep learning functionalities to large-scale big data programs and/or workflow**
- **Leverage existing Hadoop/Spark clusters to run deep learning applications**
  - **Shared, monitored and managed with other workloads (*e.g., ETL, data warehouse, feature engineering, traditional ML, graph analytics, etc.*) in a dynamic and elastic fashion**

# BigDL

## Bringing Deep Learning To Big Data Platform

- Distributed deep learning framework for Apache Spark*
- Make deep learning more accessible to big data users and data scientists
  - Write deep learning applications as ***standard Spark programs***
  - Run on existing Spark/Hadoop clusters (***no changes needed***)
- Feature parity with popular deep learning frameworks
  - E.g., Caffe, Torch, Tensorflow, etc.
- High performance (on CPU)
  - Powered by Intel MKL and multi-threaded programming
- Efficient scale-out
  - Leveraging Spark for distributed training & inference




https://github.com/intel-analytics/BigDL

https://bigdl-project.github.io/

# BigDL Run as Standard Spark Programs

## Standard Spark jobs

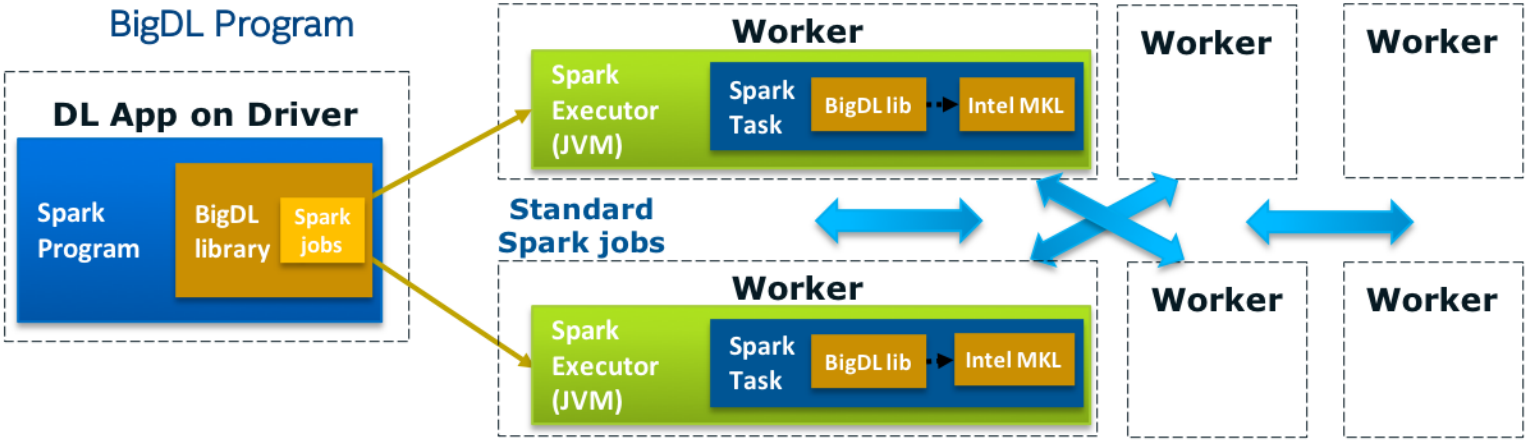
- No changes to the Spark or Hadoop clusters needed

## Iterative

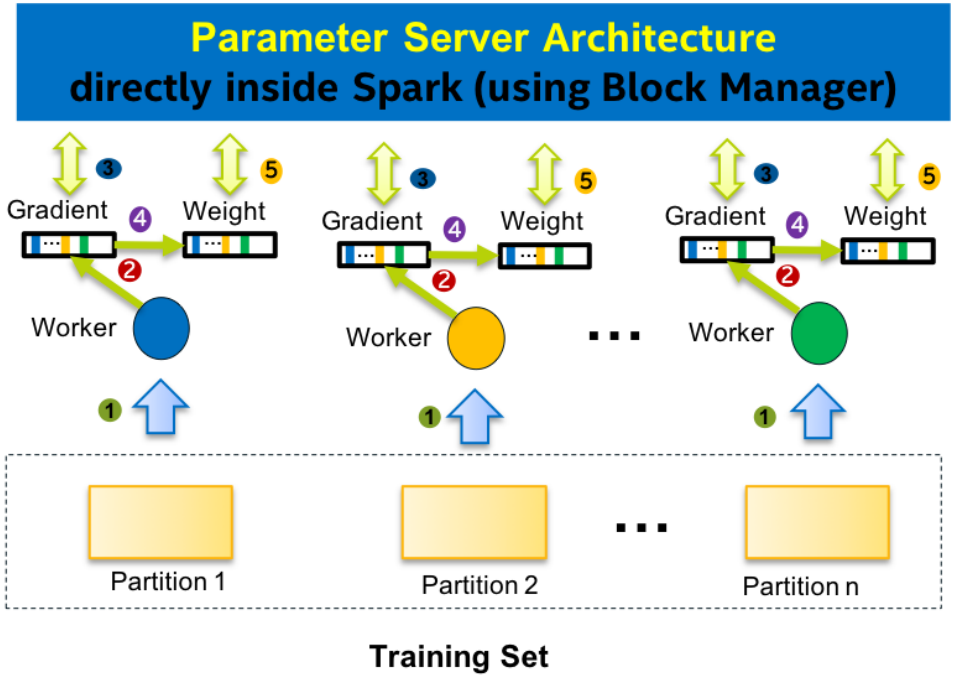
- Each iteration of the training runs as a Spark job

## Data parallel

- Each Spark task runs the same model on a subset of the data (batch)



# Distributed Training in BigDL




Peer-2-Peer All-Reduce Synchronization

# Analytics Zoo

## Unified Analytics + AI Platform for Big Data

### Distributed TensorFlow, Keras and BigDL on Spark

| | |
|---|---|
| **Reference Use Cases** | • Anomaly detection, sentiment analysis, fraud detection, image generation, chatbot, etc. |
| **Built-In Deep Learning Models** | • Image classification, object detection, text classification, text matching, recommendations, sequence-to-sequence, anomaly detection, etc. |
| **Feature Engineering** | Feature transformations for<br>• Image, text, 3D imaging, time series, speech, etc. |
| **High-Level Pipeline APIs** | • Distributed TensorFlow and Keras on Spark<br>• Native support for transfer learning, Spark DataFrame and ML Pipelines<br>• Model serving API for model serving/inference pipelines |
| **Backbends** | Spark, TensorFlow, Keras, BigDL, OpenVINO, MKL-DNN, etc. |

https://github.com/intel-analytics/analytics-zoo/ https://analytics-zoo.github.io/

# Analytics Zoo

**Build end-to-end deep learning applications for big data**

- Distributed *TensorFlow* on Spark
- *Keras*-style APIs (with autograd & transfer learning support)
- *nnframes*: native DL support for Spark DataFrames and ML Pipelines
- Built-in *feature engineering* operations for data preprocessing

**Productionize deep learning applications for big data at scale**

- *Model serving* APIs (w/ OpenVINO support)
- Support Web Services, Spark, Storm, Flink, Kafka, etc.

**Out-of-the-box solutions**

- Built-in deep learning *models* and reference *use cases*

# What Can you do with Analytic Zoo?

Anomaly Detection

- Using LSTM network to detect anomalies in time series data

Fraud Detection

- Using feed-forward neural network to detect frauds in credit card transaction data

Recommendation

- Use Analytics Zoo Recommendation API (i.e., Neural Collaborative Filtering, Wide and Deep Learning) for recommendations on data with explicit feedback.

Sentiment Analysis

- Sentiment analysis using neural network models (e.g. CNN, LSTM, GRU, Bi-LSTM)

Variational Autoencoder (VAE)

- Use VAE to generate faces and digital numbers

https://github.com/intel-analytics/analytics-zoo/tree/master/apps

# Building and Deploying with BigDL/Analytics Zoo


http://software.intel.com/bigdl/build

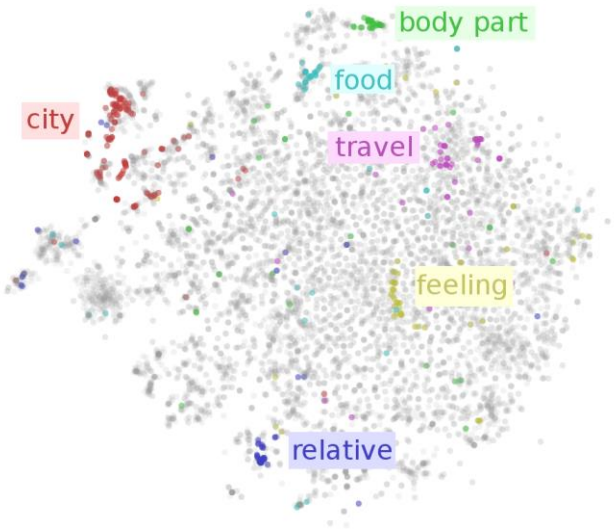
*Not a Full List*

# Agenda

- Job/Resume Search Challenges and Opportunity
- Analytics Zoo and BigDL Overview
- **Resume Search Analytics Zoo Solution**
- Takeaways

# Word Embeddings and GloVe Vectors

- **Words or phrases from the vocabulary are mapped to vectors of real numbers.**
- **Global log-bilinear regression model for the unsupervised learning algorithm.**
- **Training is performed on aggregated global word-word co-occurrence statistics from a Wikipedia.**
- **Vector representations showcase meaningful linear substructures of the word vector space.**

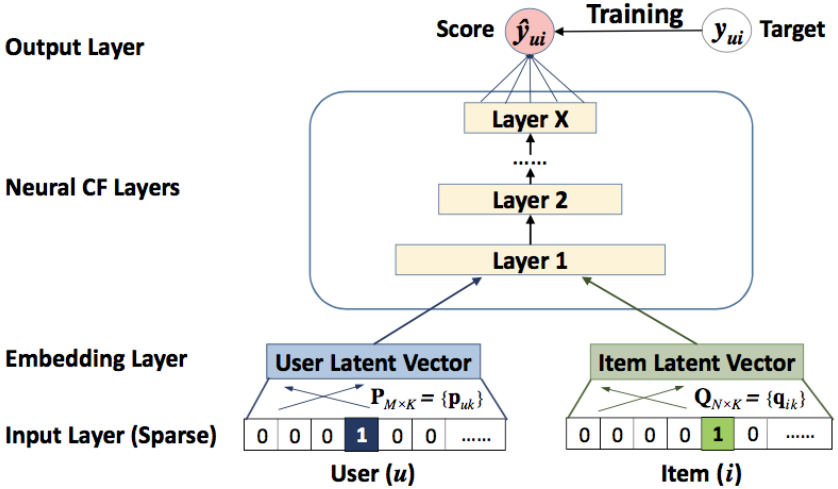





https://nlp.stanford.edu/projects/glove/

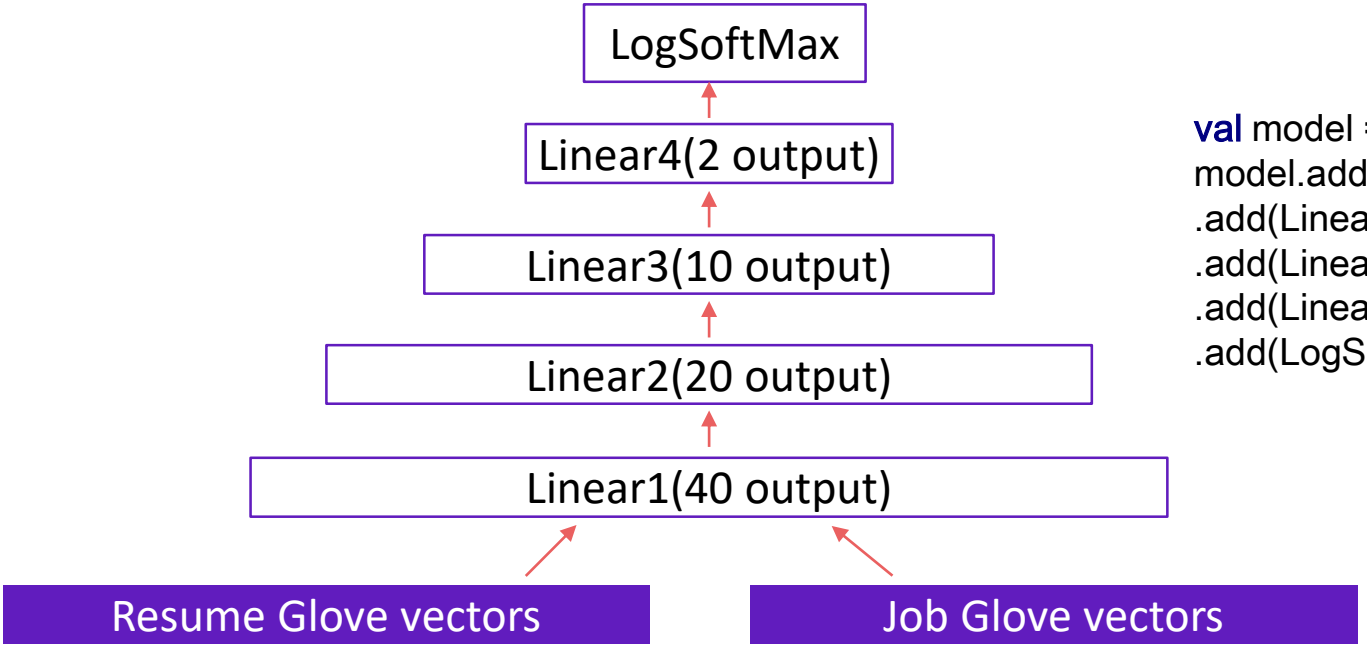
# Analytics Zoo Recommender Model

- Neural collaborative filtering, Wide and Deep
- Answer the question using classification methodologies
- Implicit feedback and explicit feedback
- APIs
  - recommendForUser
  - recommendForItem
  - predictUserItemPair



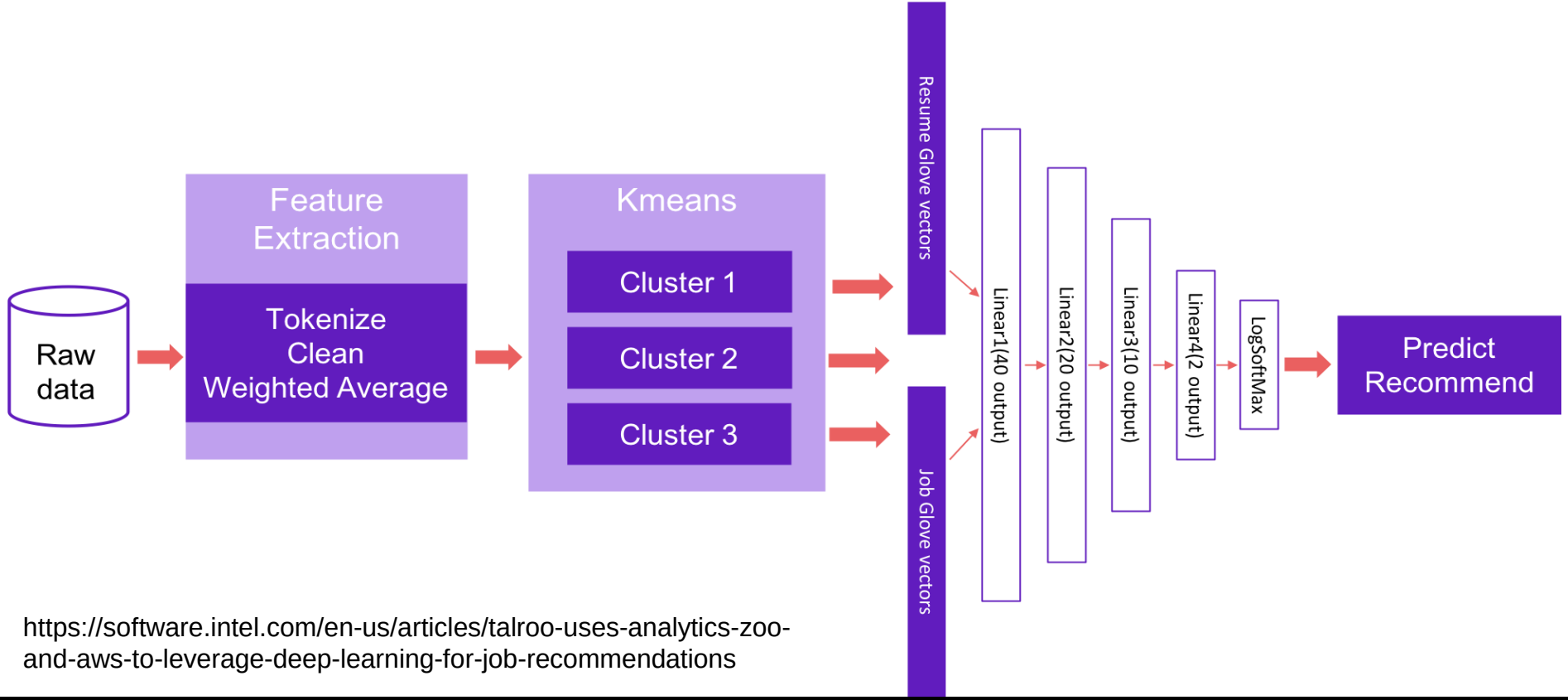

He, 2015

# Recommender model



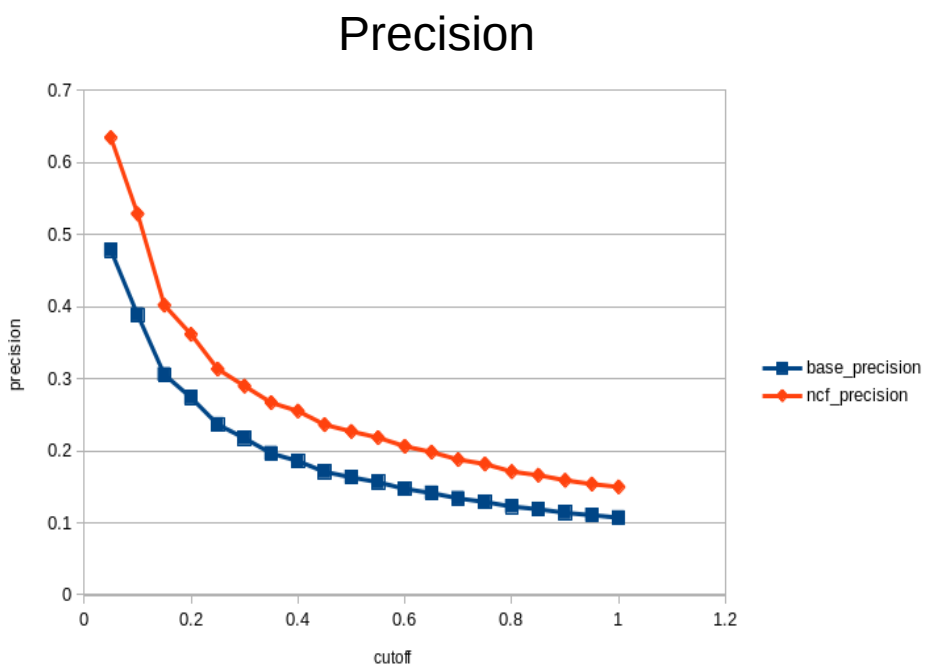

```
val model = Sequential[Float]()
model.add(Linear(100, 40)).add(ReLU())
.add(Linear(40, 20)).add(ReLU())
.add(Linear(20, 10)).add(ReLU())
.add(Linear(10, 2)).add(ReLU())
.add(LogSoftMax())
```

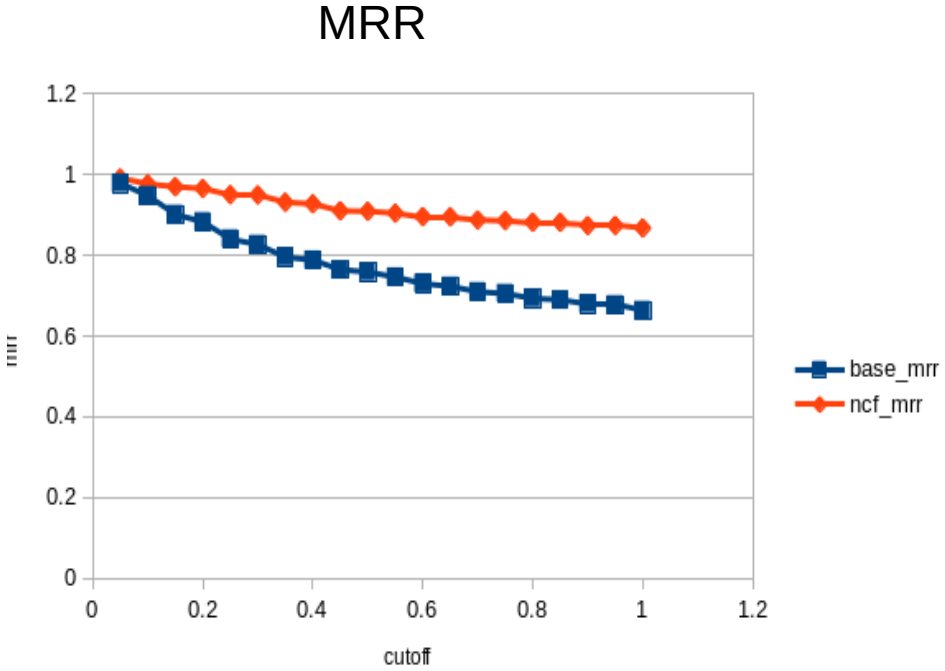
# End to End Flow




https://software.intel.com/en-us/articles/talroo-uses-analytics-zoo-and-aws-to-leverage-deep-learning-for-job-recommendations

# Evaluation Results



# Takeaways

- Analytics Zoo/BigDL integrates well into existing AWS Databricks Spark ETL and machine learning platform
- Analytics Zoo/BigDL scales with our data and business
- Jobs and resumes can be effectively modeled and processed through embeddings
- Ensembling multiple models and glove embedding feature embedding proved to be very effective for rich content
- More information available at https://analytics-zoo.github.io/

# ANALYTICS ZOO

## Unified Analytics + AI Platform

Distributed TensorFlow, Keras and BigDL on Apache Spark

https://github.com/intel-analytics/analytics-zoo

# Legal Disclaimers

- Intel technologies' features and benefits depend on system configuration and may require enabled hardware, software or service activation. Learn more at intel.com, or from the OEM or retailer.
- No computer system can be absolutely secure.
- Tests document performance of components on a particular test, in specific systems. Differences in hardware, software, or configuration will affect actual performance. Consult other sources of information to evaluate performance as you consider your purchase. For more complete information about performance and benchmark results, visit **http://www.intel.com/performance**.

Intel, the Intel logo, Xeon, Xeon phi, Lake Crest, etc. are trademarks of Intel Corporation in the U.S. and/or other countries.
*Other names and brands may be claimed as the property of others.
© 2019 Intel Corporation