FINAL YEAR PROJECT

2019 - 2020

DEVAL MUDIA (BT16CSE060)
AYUSH SINGH (BT16CSE098)
CHAITYA CHHEDA (BT16CSE016)
SADNEYA PUSALKAR (BT16CSE076)

UNDER THE GUIDANCE OF
Dr. S.R. SATHE

PROBLEM STATEMENT

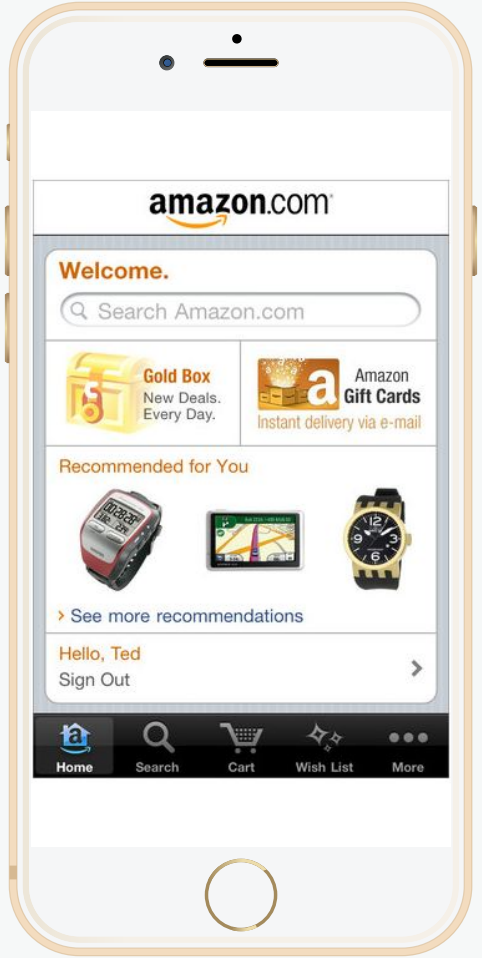
Parallelized Recommendation System using Spark and OpenMP

PROBLEM DESCRIPTION

Companies like Amazon and Netflix rely on targeted recommendations in order to serve a broad range of products/content to its users.

These recommendations are based on a user's previous history as well as products/content that similar users have purchased/watched.

Therefore, computing how similar two users are is an essential part of the recommendation process.



BIG DATA

Amazon and Netflix dataset is not neatly organised into a matrix of users and products.

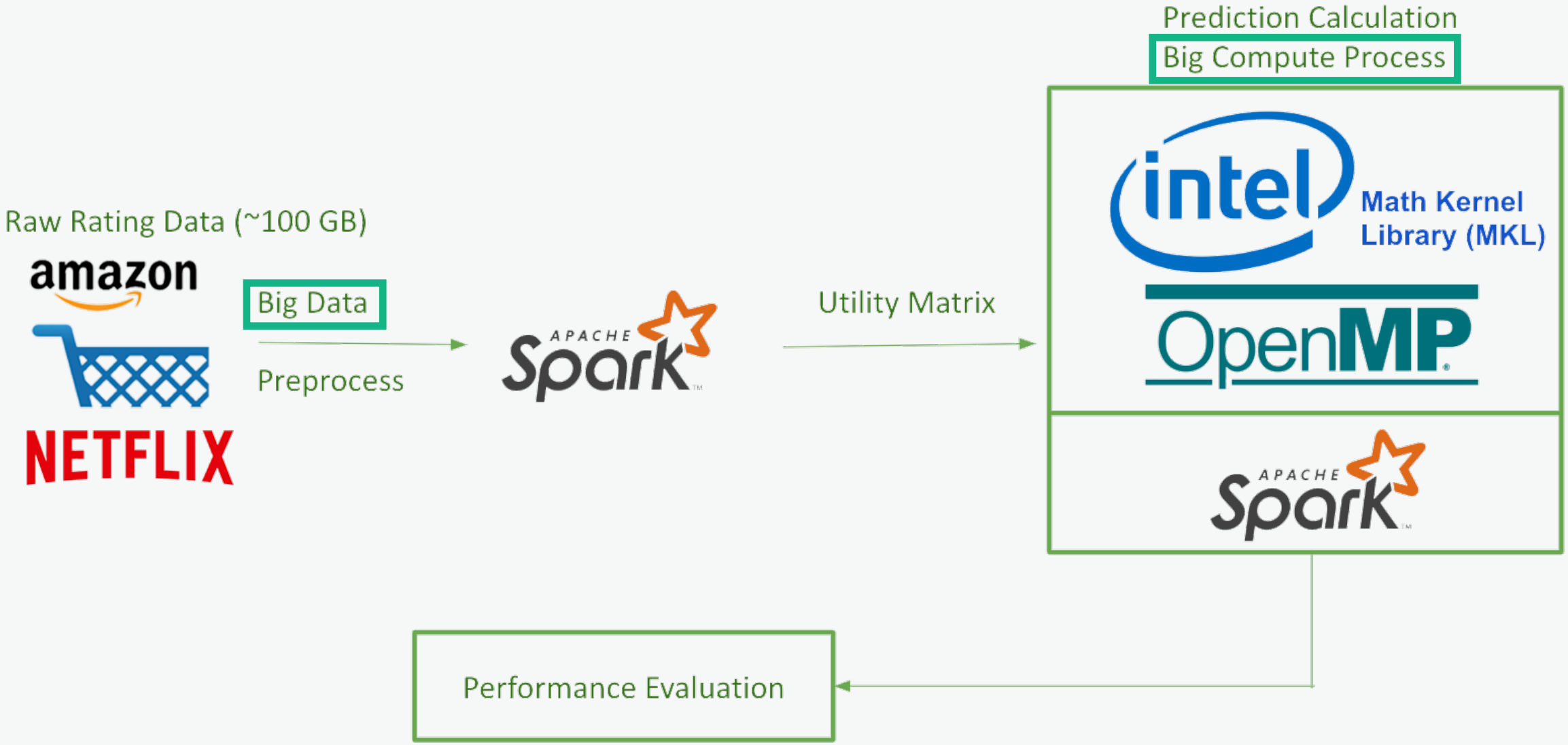
We are dealing with a large, unstructured dataset and in order to process it into a matrix of this form, we would need to make use of **big data processing** solutions such as Spark.

BIG COMPUTE

In order to compute similarity scores and generate predictions, we rely on a lot of matrix or vector products.

These matrix operations can be made parallel through **big compute** using multi-threading to speed up these computations.

Platform and Infrastructure



WHY Apache Spark

Apache Spark is a general-purpose & lightning fast cluster computing system.

Spark is 100 times faster
than Bigdata Hadoop and 10 times faster
than accessing data from disk.

WHY OpenMP

Enabling HPC since 1997

Ability to **parallelise** small parts of an application at a time.

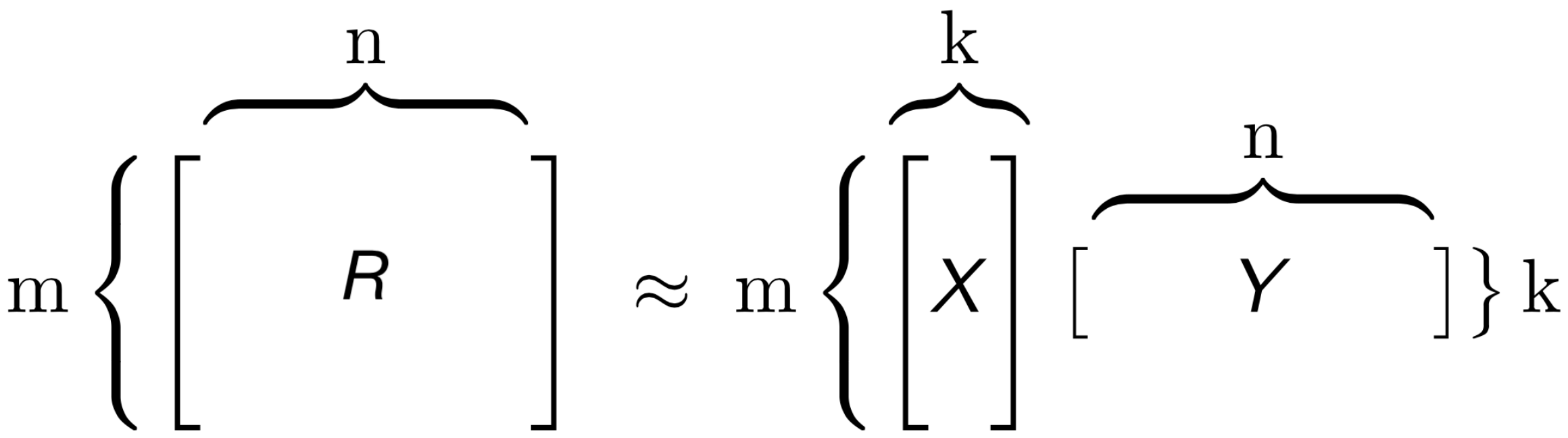
Impact on code quantity and quality(**readability**).

Availability of application development and debugging environments.

Recommendation System Model

Content-based systems : Aims to assess the features of the products being bought.

Collaborative filtering systems : Aims to assess the users purchasing items.



Raw JSON Data


```
{
  "reviewerID": "A2SUAM1J3GNN3B",
  "asin": "0000013714",
  "reviewerName": "J. McDonald",
  "helpful": [2, 3],
  "reviewText": "It looks good and solve my problems",
  "overall": 5.0,
  "summary": "Heavenly Highway Hymns",
  "unixReviewTime": 1252800000,
  "reviewTime": "09 13, 2009"
}
```


Model

	P1	P2	P3	P4	P5	P6
U1	3		4	5		
U2	1				5	
U3			4			3
U4		5	2		5	
U5	3			5		4

Goal

	P1	P2	P3	P4	P5	P6
U1	3	?	4	5	?	?
U2	1	?	?	?	5	?
U3	?	?	4	?	?	3
U4	?	5	2	?	5	?
U5	3	?	?	5	?	4

SEMESTER PLAN

Understanding Parallel Computing

Hardware Interface Algorithmic Concepts

OpenMP

Parallelizing Sample Algorithms

Understanding Big Data

Apache Spark

Understanding Recommendation System Models

Standard Collaborative Filtering Model (SCF)

Matrix Factorisation (MF)

Alternative Least Square (ALS)

Resources

An Introduction to OpenMP

Intel Corp.

by Tim Mattson

High Performance Computing

Udacity Course

Georgia Tech CS6220

Extreme Scale Data and Computational Science

Harvard-CS

Links

https://classroom.udacity.com/courses/ud281

[1] Jeffrey D. Ullman, “Mining Massive Datasets: Recommendation Systems”[Online]. Available: http://infolab.stanford.edu/~ullman/mmds/ch9.pdf

[2] Haoming Li, Bangzheng He, Michael Lublin, Yonathan Perez, “Matrix Completion via Alternating Least Square(ALS)” [Online]. Available: http://stanford.edu/~rezab/classes/cme323/S15/notes/lec14.pdf

THANK YOU