CSC 591 – Data Intensive Computing Fall ‘17

MapReduce performance evaluation on EC2 (HDFS HA) vs EMR

Application - Online spam reviewers detection

Instructor
Dr. Vincent Freeh

Team 3
Divya Guttikonda
Nithya Kumar
Sahithi Guddeti

Agenda

- ✓ Problem Statement
- ✓ Solution Overview
- ✓ Application Workflow
- ✓ MapReduce Phase
- ✓ HDFS High Availability Architecture on Amazon EC2
- ✓ Amazon EMR Architecture and Workflow
- ✓ Performance Evaluation
- ✓ Conclusion

Problem Statement

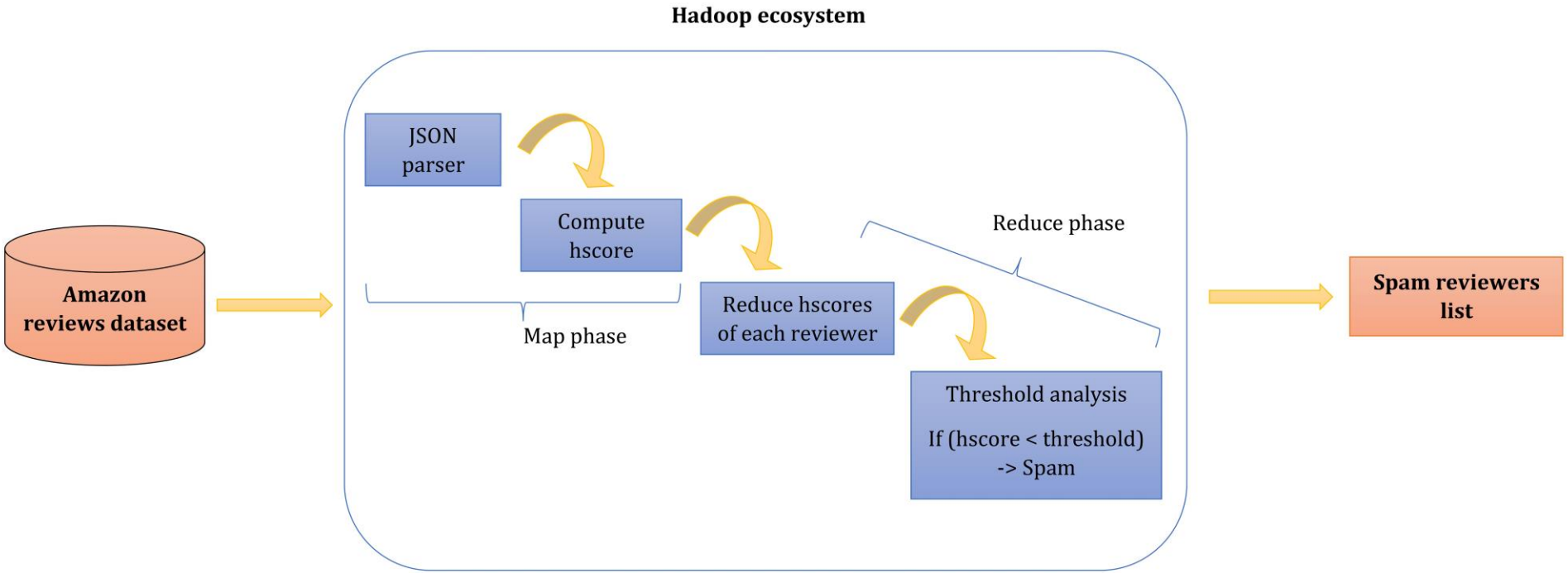
- ✓ Opinion spamming in e-commerce reviews
- ✓ Analyzing and handling huge volumes of data

Solution Overview

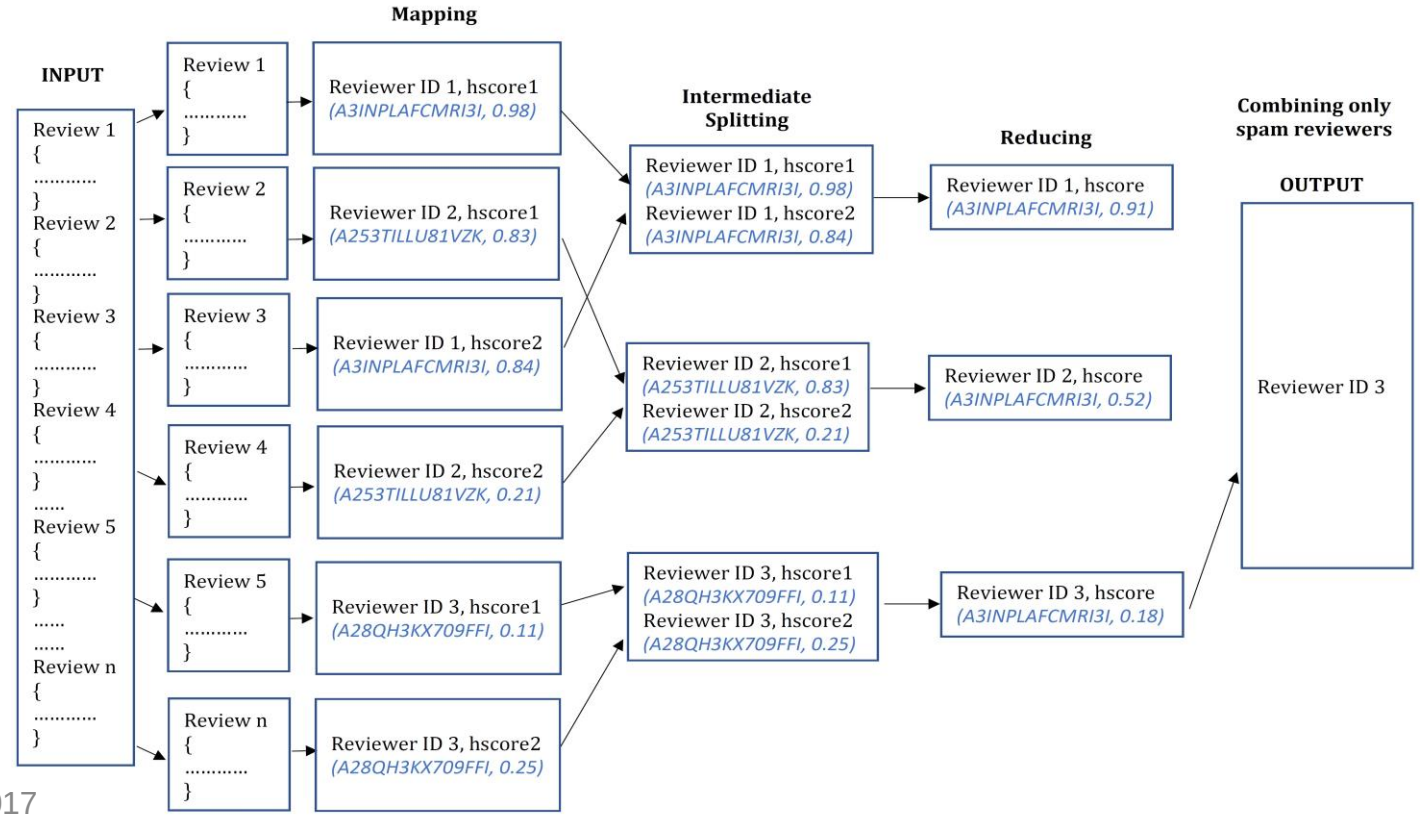
- ✓ MapReduce framework to detect online spam reviewers

- ✓ Infrastructure evaluation
  - Hadoop HDFS cluster with HA (EBS storage) and EMR cluster (S3 storage)

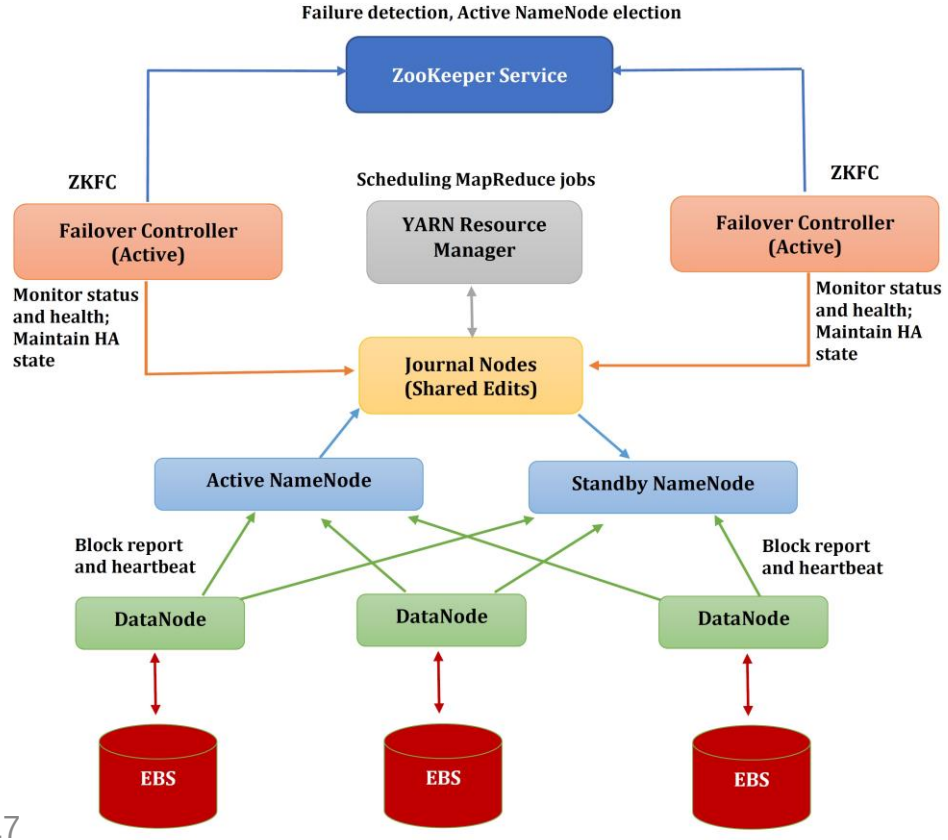
Application Workflow

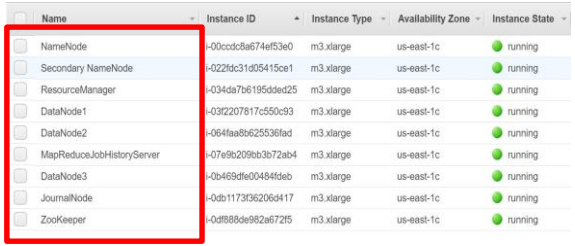


MapReduce Phase



HDFS High Availability Architecture on Amazon EC2

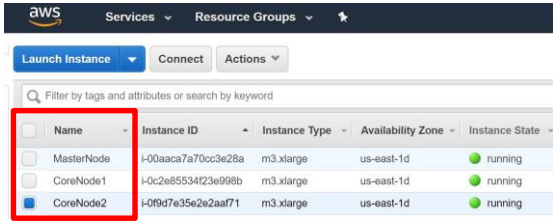




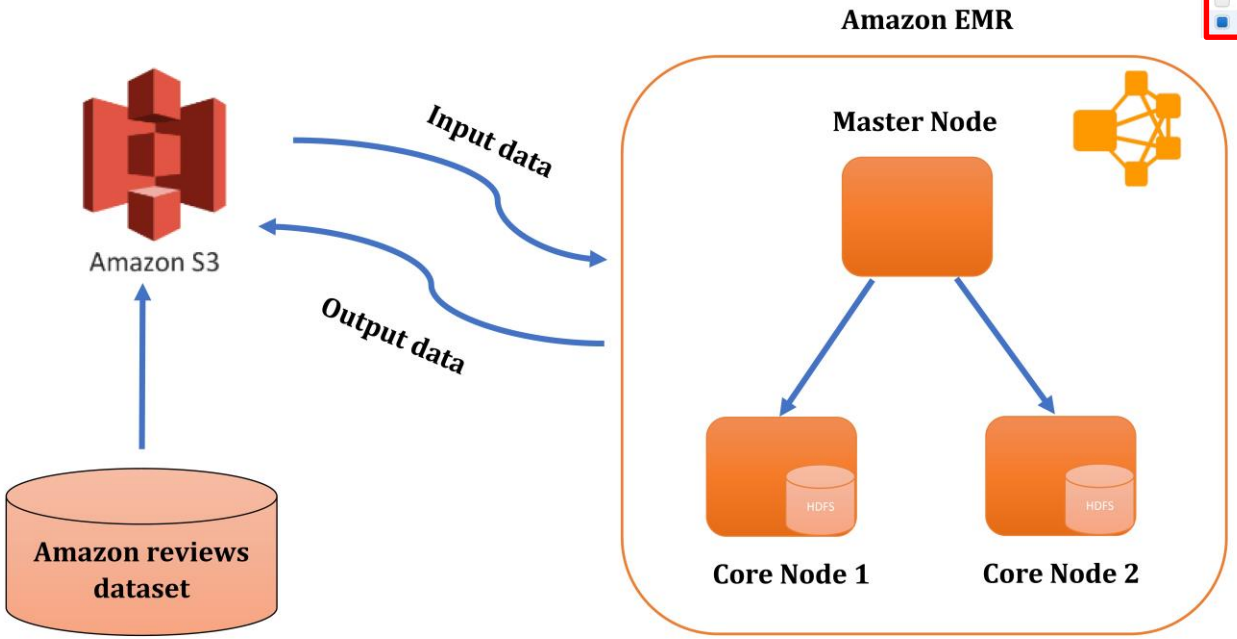
Name	Instance ID	Instance Type	Availability Zone	Instance State
NameNode	i-00ccdc8a674ef53e0	m3.xlarge	us-east-1c	running
Secondary NameNode	i-022fdc31d05415ce1	m3.xlarge	us-east-1c	running
ResourceManager	i-034da7b6195dded25	m3.xlarge	us-east-1c	running
DataNode1	i-03f2207817c550c93	m3.xlarge	us-east-1c	running
DataNode2	i-064faa8b625536fad	m3.xlarge	us-east-1c	running
MapReduceJobHistoryServer	i-07e9b209bb3b72ab4	m3.xlarge	us-east-1c	running
DataNode3	i-0b469dfe00484fdeb	m3.xlarge	us-east-1c	running
JournalNode	i-0db1173f36206d417	m3.xlarge	us-east-1c	running
ZooKeeper	i-0df888de982a672f5	m3.xlarge	us-east-1c	running

Running instances on EC2 for HDFS HA

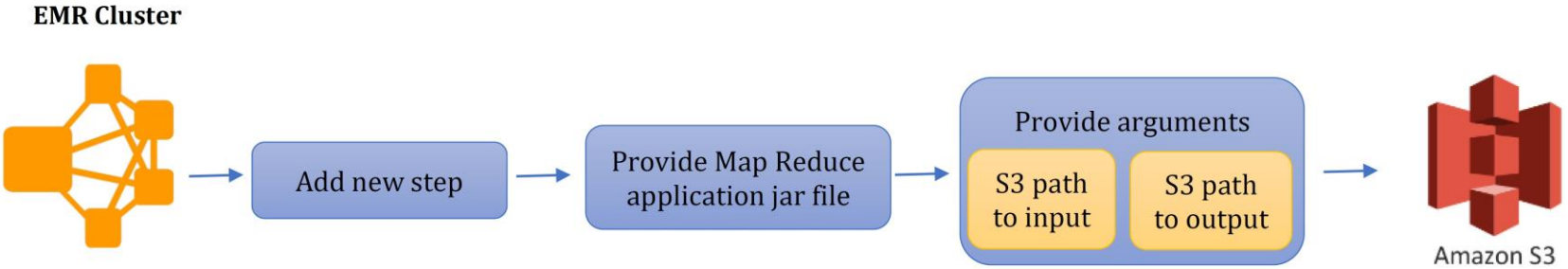
Amazon EMR Architecture



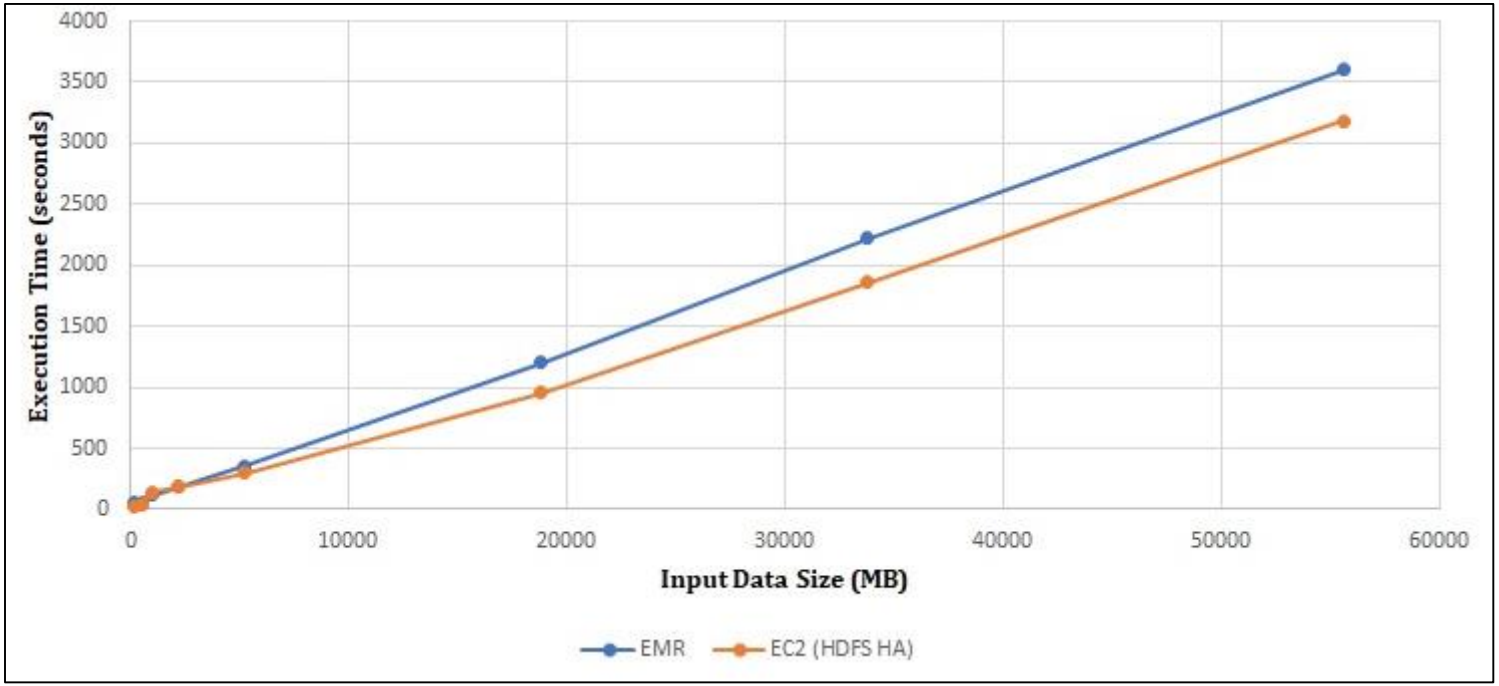
Running instances on EC2 for EMR



EMR Workflow



Performance Evaluation



Conclusion

- ✓ MapReduce application for reviewers detection → Optimal, cost-effective
- ✓ HDFS High Availability cluster → High consistency and storage near to compute
- ✓ EMR cluster on S3 → Scalable, simple and flexible

Questions?!?