SPARK OVER HADOOP(The Distributed programming framework)

Ankita K. Wani

SSBT'S COET,Bambhori,Jalgaon

October 8, 2015

outline

Introduction

- **Spark is an open-source distributed programming framework .**
- **Spark is a cluster framework that performs in-memory computing, with the goal of outperforming disk-based engines like Hadoop .**
- **Supports advanced Python, Scala ,SparkR ,Java8, SQL.**
- **Spark is a data parallel open source processing framework.**
- **An alternative to the traditional batch map/reduce model.**
- **It is used for real-time stream data processing and fast interactive queries .**

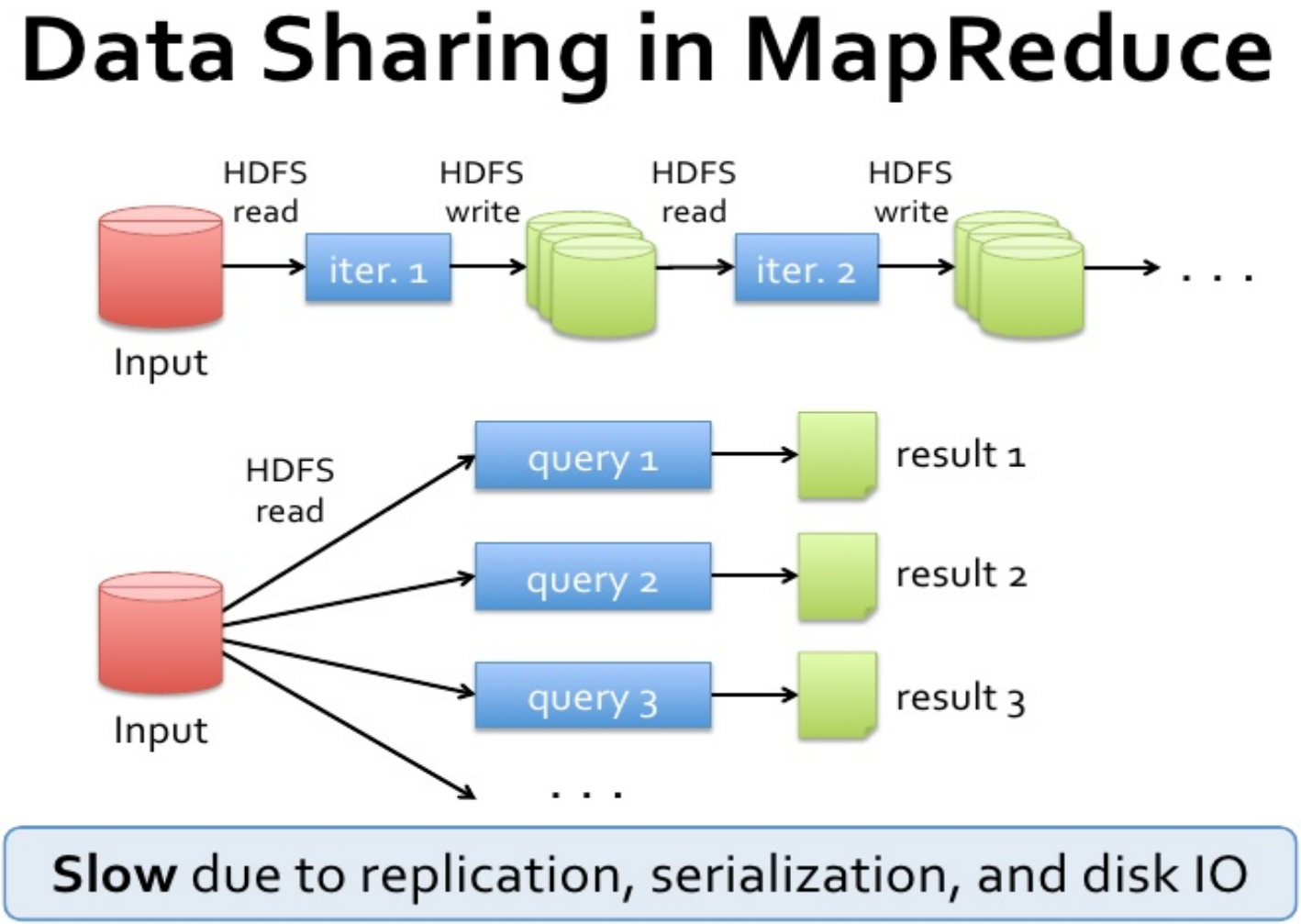
Features

- **In-memory (RAM) based**
- **Real-time stream data processing**
- **Makes big data interactive.**
- **Parallel open source processing.**
- **2-5 times less code.**
- **Scalability to over 8000 nodes in production.**
- **Support for structured and relational query processing (SQL).**
- **Easy Management.**

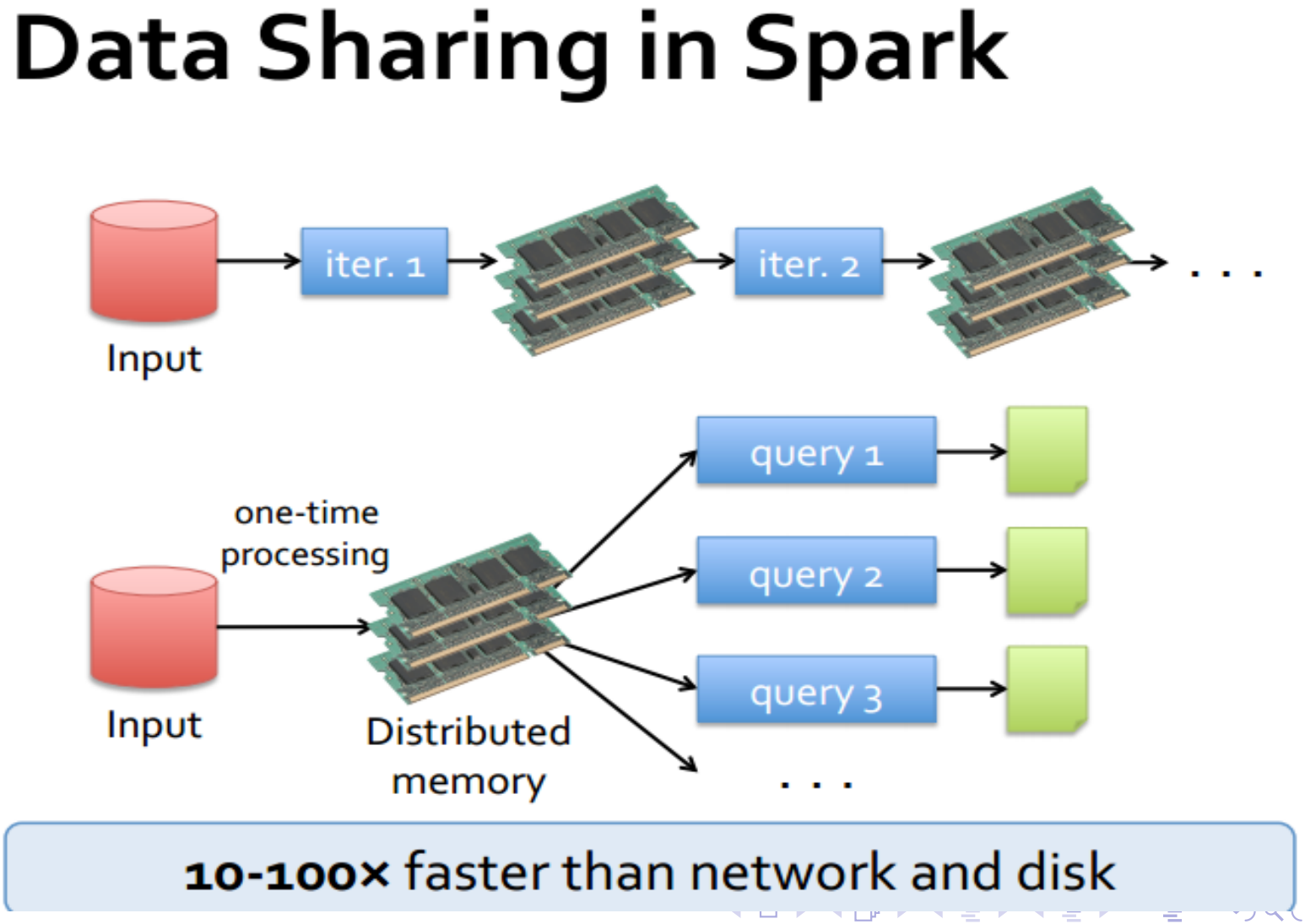
History

- **Originally developed by Matei Zaharia in 2009 at UC Berkeley's AMP Lab written in Scala**
- **Released as open source in 2010**
- **In 2013, It was donated to Apache Software Foundation.**
- **In February 2014, Spark became a Top-Level Apache Project.**

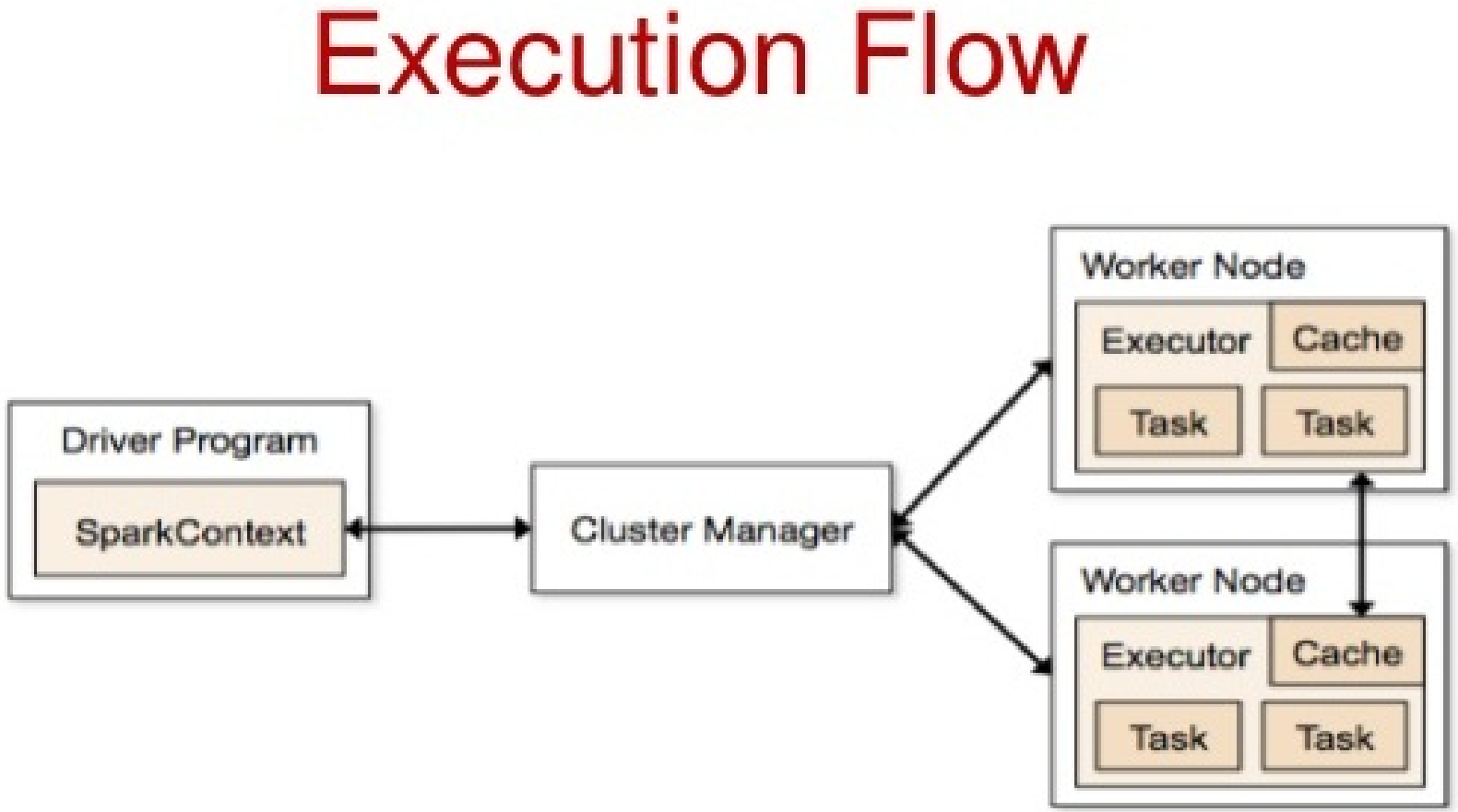
Data sharing



Data sharing



Execution


Execution Flow
Worker Node
Executor
Cache
Task
Task
Driver Program
SparkContext
Cluster Manager
Worker Node
Executor
Cache
Task
Task

Related Terms

- Cluster Manager
- Executer
- Driver
- Stage

Applications

- In-memory Analytics and anamoly detection.
- Traffic Estimation.
- Twitter scam classification.

Disadvantages

- Little bit less secure than Hadoop. -RDD creates some problem in security.
- In case fault tolerance it will have to start processing from begining.

Conclusion

By making duistributed datasets a first-class primitives, spark provides a simple, efficient programming model for stateful data analytics.
RDD supports in-memory catching, fault recovery and debugging.


Thank You