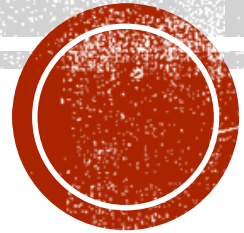
BUILDING A HIGH LEVEL DATAFLOW SYSTEM ON TOP OF MAP-REDUCE



Presentation (paper) By: Michael Sanzo

Date: October 20, 2016

Professor: Alan Labouseur

QUICK OVERVIEW

"The Pig Experience"
- Slide 3- Main Ideas
- Slide 4- How the Idea was Implemented
- Slide 5- Analysis

"Large Scale Data Analysis"
- Slide 6- Main Ideas
- Slide 7- How the Idea was Implemented
- Slide 8- Analysis
- Slide 9- Comparison

"Michael Stonbraker"
- Slide 10- Main Ideas

Summary
- Slide 11- Advantages/Disadvantages

MAIN IDEA - PIG

Initially developed at Yahoo!

Focus on what they want to do, rather than how it gets done.

Main Idea: Pig programming language is designed to handle any kind of data with much more efficiency

2 components

- Pig Latin
- Pig Runtime Environment

HOW IDEA WAS IMPLEMENTED - PIG

- Project Experience
- 2 Components
  - The Programming Language. How it is much easier than having to write mapper and reduce programs.
    - Load
    - Transformations
    - DUMP or STORE
  - Runtime Environment
- Streaming
- User Defined Functions (UDFs)

ANALYSIS - PIG

- Open Source
- Pig Mix Benchmark
- Data Generator
- Results

MAIN IDEAS - LSDP

- New computing model
- Two Approaches to Large Scale Data Analysis
  - MapReduce
  - Parallel DBMSs
- Architectural Elements
  - Schema Support
  - Indexing
  - Modeling
  - Data Distribution Strategy
  - Flexibility
  - Fault Tolerance

HOW IDEA WAS IMPLEMENTED - LSDP

- Hadoop
- DBMS-X
- Vertica

Analytical Tasks

- Data Loading
- Selection Task
- Aggregation Task
- Join Task
- UDF Aggregation Task

ANALYSIS - LSDP

- This performance advantage that the two database systems share is the result of a number of technologies that had been developed over the past 25 years.

B-tree indices

Novel storage mechanisms

Compression Techniques

Sophisticated Parallel Algorithms

COMPARISON – PIG AND LSDP

- Lessen the burden of implementing repetitive tasks
- Pig is a higher level interface
- Complex tasks easier to code

MAIN IDEAS - STONEBRAKER

- Data Warehouse Market
- OLTP Market
- NoSQL Market
- Complex Analytics
- Streaming Market
- Graph Analytics Market

ADVANTAGES AND DISADVANTAGES

Advantages

- Pig Latin is Procedural
- Check pointing Data
- Faith in Optimizers
- User Interaction/Inserting Developer Codes
- DAGs

Disadvantages

- Data Schema
- Diversity of Programs