Daugherty

BUSINESS SOLUTIONS

# Hadoop Data Modeling

**05-02-2018**

# Agenda

- What's the Big Data Innovation, Data Engineering, Analytics Group?
- Data Modelling in Hadoop
- Questions

# It started with an article

datanami

DATA SCIENCE • AI • ADVANCED ANALYTICS

Home About Whitepapers Events Subscribe

HOME | FEATURES | SECTORS | APPLICATIONS | TECHNOLOGIES | VENDORS

Top Stories On

September 29, 2017

## Hadoop Was Hard to Find at Strata This Week

Alex Woodie

It's been barely six months since the word "Hadoop" was removed from the name of the world's biggest big-data trade show, and at this week's Strata Data Conference in New York City, Hadoop all but disappeared. Yet key parts of the Hadoop platform appear likely to survive.

It was an auspicious absence, to be sure. Making a big yellow elephant essentially vanish in the space of half a year is not an easy feat. But the fact remains that what used to be the rallying point for an entire industry has essentially been reduced to an afterthought. Cloudera, which puts on the show with O'Reilly Media, scarcely even mentioned Hadoop.

## And a name change

# Data Eng Weekly

## Your weekly Data Engineering news

*Formerly Hadoop Weekly*

Data Eng Weekly covers the week's top news in the data engineering ecosystem. Each issue (delivered on Sunday), keeps subscribers up to date on the latest data engineering-related open source and cloud news across batch (e.g. Apache Hadoop, Apache Spark), stream (e.g. Apache Kafka), distributed systems, and much more.

# Which led to some questions

- What is the future of the Hadoop ecosystem?
- What is the dividing line between Spark and Hadoop?
- What are the big players doing?
- How does the push to cloud technologies affect Hadoop usage?
- How does Streaming come into play?

# And then our answer

- Hadoop is here to stay, but it will make the most strides as a machine learning platform.
- Spark can perform many of the same tasks that elements of the Hadoop ecosystem can, but it is missing some existing features out of the box.
- Cloudera, Hortonworks, and MapR are positioning themselves as data processing platforms with roots in Hadoop, but other aspirations. For example, Cloudera is positioning itself as a machine learning platform.
- The push to cloud means that the distributed filesystem of HDFS may be less important to cloud-based deployments. But Hadoop ecosystem projects are adapting to be able to work with cloud sources.
- The Hadoop ecosystem projects have proven patterns for ingesting streaming data and turning it into information.

# Introducing …

- We’re now going to be

St. Louis Big Data Innovation, Data Engineering, and Analytics Group

Or more simply put:

St. Louis Big Data IDEA

# So What is the STL Big Data IDEA interested in?

- Local Companies
- Big Data
  - Hadoop
  - Cloud deployments
  - Cloud-native technologies
  - Spark
  - Kafka
- Innovation
  - New Big Data projects
  - New Big Data services
  - New Big Data applications
- Data Engineering
  - Streaming data
  - Batch data analysis
  - Machine Learning Pipelines
  - Data Governance
  - ETL @ Scale
- Analytics
  - Visualization
  - Machine Learning
  - Reporting
  - Forecasting

# Introducing our New Board Member

- Scott Shaw has been with Hortonworks for four years.
- He is the author of four books including Practical Hive and Internet of Things and Data Analytics Handbook.
- Scott will be helping our group find speakers in the open source community.

Please help me welcome Scott to the group in his new role

# Agenda

- The Schema-on-Read Promise
- File formats and Compression formats
- Schema Design – Data Layout
- Indexes, Partitioning and Bucketing
- Join Performance
- Hadoop SQL Boost – Tez, Cost Based Optimizations & LLAP
- Summary

# Introducing our Speakers

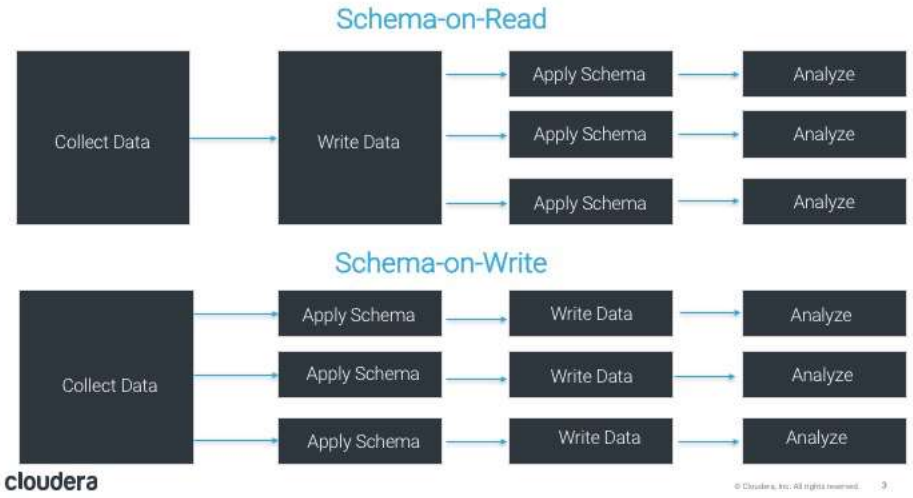
## Adam Doyle

- Co-Organizer, St. Louis Big Data IDEA
- Big Data Community Lead, Daugherty Business Solutions

## Drew Marco

- Board Member & Secretary, TDWI
- Data and Analytics Line of Service Leader, Daugherty Business Solutions

# Schema On Read




## Schema on Write

- Schemas are typically purpose-built and hard to change
- Generally loses the raw/atomic data as a source
- Requires considerable modeling/implementation effort before being able to work with the data
- If a certain type of data can't be confined in the schema, you can't effectively store or use it (if you can store it at all)

## Schema on Read

- Slower Results
- Preserve the raw/atomic data as a source
- Flexibility to add, remove and modify columns
- Data may be riddled with missing or invalid data, duplicates
- Suited for data exploration and not recommended for repetitive querying and high performance

**Real world use of Hadoop / Hive that require high performing queries on large data sets requires up-front planning and data modeling**

# Schema Design – Data Layout

*"The primary reason to avoid normalization is to minimize disk seeks, such as those typically required to navigate foreign key relations. Denormalizing data permits it to be scanned from or written to large, contiguous sections of disk drives, which optimizes I/O performance. However, you pay the penalty of denormalization, data duplication and the greater risk of inconsistent data."*

Source: *Programming Hive* by Dean Wampler, Jason Rutherglen, Edward Capriolo, O'Reilly Media

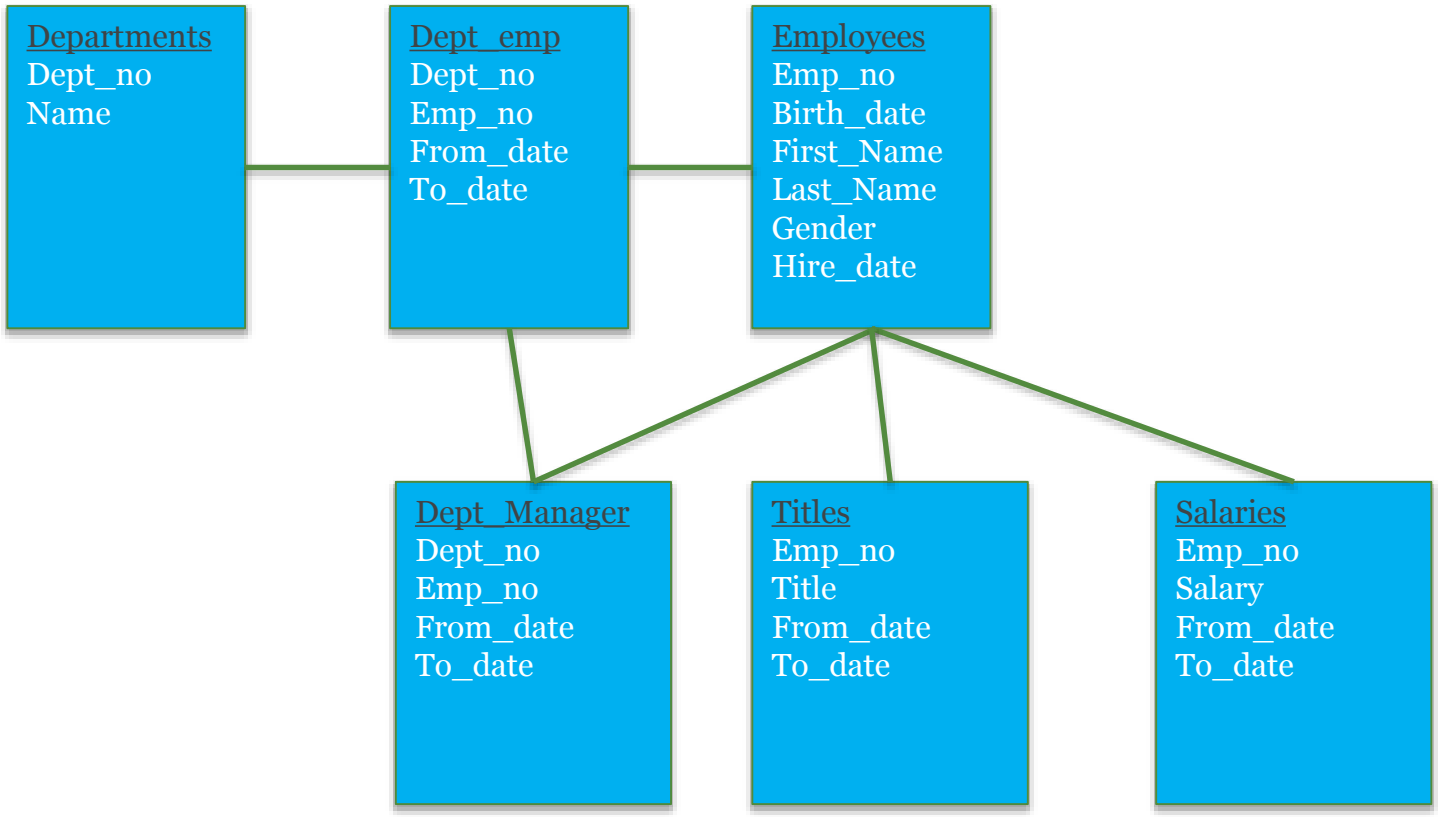
## *Normalization*

- Pros
  - Reduces data redundancy
  - Decreases risk of inconsistent datasets
- Cons
  - Requires re-organization of source data
  - Less efficient storage

## *Denormalization*

- Pros
  - Often requires reorganizing the data (slower writes)
  - Minimizes disk seeks (i.e. FK relations)
  - Storage in large contiguous disk drive segments
- Cons
  - Data Duplication
  - Increased Risk of inconsistent data

# Introducing Our Use Case




https://dev.mysql.com/doc/employee/en/

# Data Storage Decisions

- Hadoop is a file system - No Standard data storage format in Hadoop
- Optimal storage of data is determined by how the data will be processed
- Typical input data is in JSON, XML or CSV

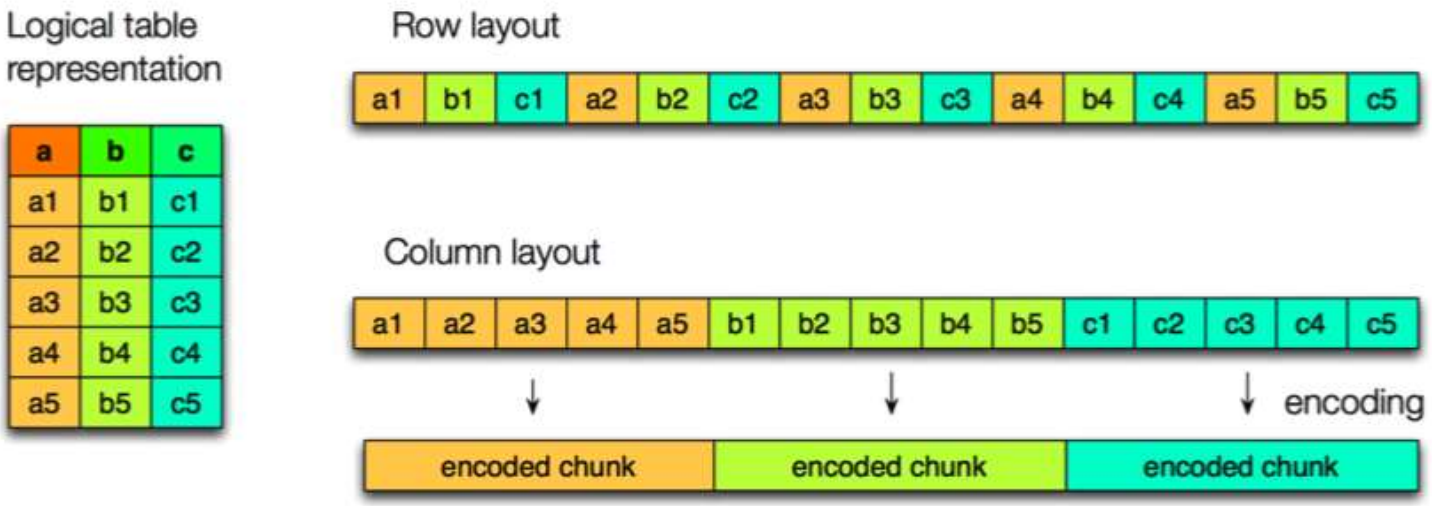
Major Considerations:

File Formats

Compression

# Parquet

- Faster access to data
- Efficient columnar compression
- Effective for select queries



# ORCFile

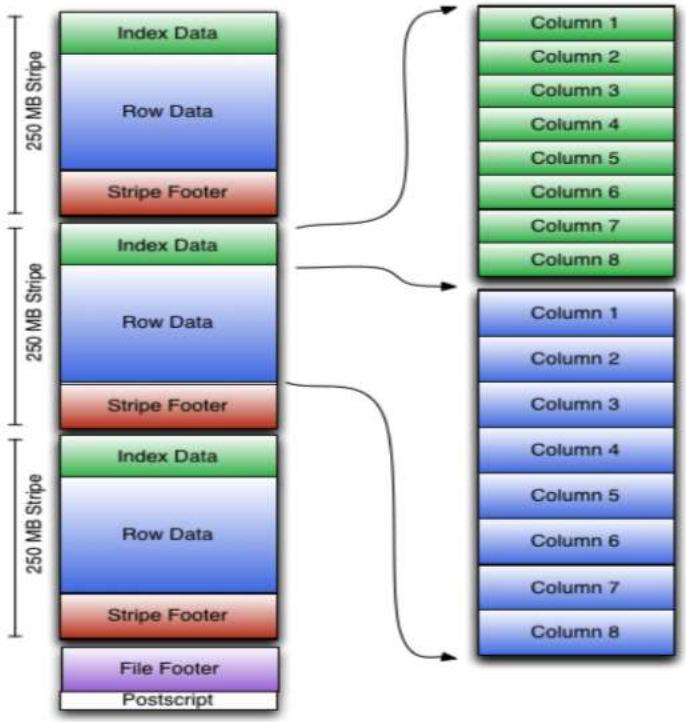
**High Performance:** Split-able, columnar storage file **Efficient Reads:** Break into large “stripes” of data for efficient read

**Fast Filtering:** Built in index, min/max, metadata for fast filtering blocks - bloom filters if desired

**Efficient Compression:** Decompose complex row types into primitives: massive compression and efficient comparisons for filtering

**Precomputation:** Built in aggregates per block (min, max, count, sum, etc.)

**Proven at 300 PB scale:** Facebook uses ORC for their 300 PB Hive Warehouse



# Avro

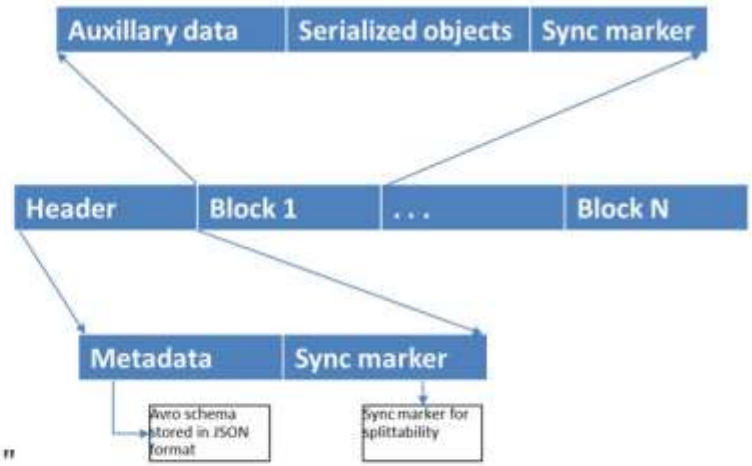
- JSON based schema
- Cross-language file format for Hadoop
- Schema evolution was primary goal – Good for Select * queries
- Schema segregated from data
- Row major format

## Avro – File structure and example

Sample AVRO schema in JSON format

```
{
  "type" : "record",
  "name" : "tweets",
  "fields" : [ {
    "name" : "username",
    "type" : "string",
}, {
    "name" : "tweet",
    "type" : "string",
}, {
    "name" : "timestamp",
    "type" : "long",
} ],
  "doc:" : "schema for storing tweets"
}
```

Avro file structure



# Comparison of file formats

| Query | Text | Avro | ORC | Parquet |
|---|---|---|---|---|
| select count(*) from employees e join salaries s on s.emp_no = e.emp_no join titles t on t.emp_no = e.emp_no; | 42.696 | 48.934 | 25.846 | 26.081 |
| select d.name, count(1), d.first_name, d.last_name from (select d.dept_no, d.dept_name as name, m.first_name as first_name, m.last_name as last_name from departments d join dept_manager dm on dm.dept_no = d.dept_no join employees m on dm.emp_no = m.emp_no where dm.to_date='9999-01-01') d join dept_emp de on de.dept_no = d.dept_no join employees e on de.emp_no = e.emp_no group by d.name, d.first_name, d.last_name; | 59.536 | 63.08 | 27.954 | 26.073 |
| Size | 124M | 134M | 16.7M | 30.5M |

# Compression

- Not just for storage (data-at-rest) but also critical for disk/network I/O (data-in-motion)
- Splittability of the compression codec is an important consideration

## Snappy

- High speed with reasonable compression
- Not splittable – only used with Avro

## LZO

- Optimized for speed as opposed to size
- Splittable but requires additional indexing
- Not shipped with Hadoop

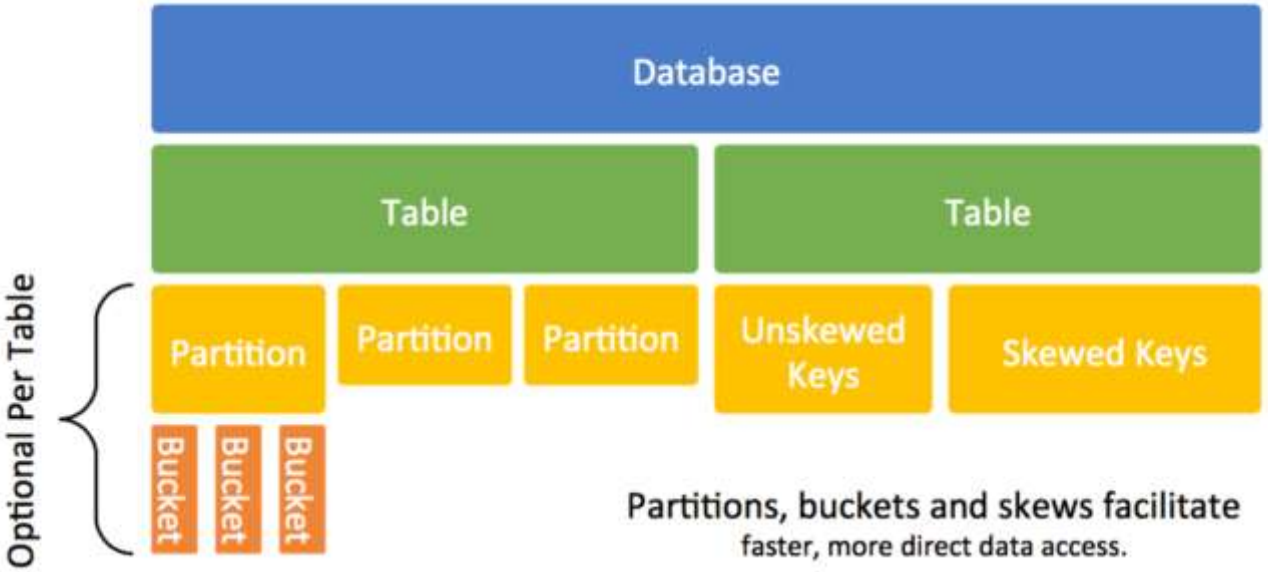
## Gzip

- Optimized for size
- Write performance is half of snappy
- Read performance as good as snappy
- Smaller blocks = better performance

## bzip2

- Optimized for size (9% better compared to Gzip)
- Splittable
- Performance sucks; Primary use is archival on Hadoop

# Partitioning & Bucketing




- Partitioning is useful for chronological columns that don't have a very high number of possible values
- Bucketing is most useful for tables that are "most often" joined together on the same key
- Skews useful when one or two column values dominate the table

# Partitioning

- Every query reads the entire table even when processing subset of data (full-table scan)
- Breaks up data horizontally by column value sets
- When partitioning you will use 1 or more “virtual” columns break up data
- Virtual columns cause directories to be created in HDFS.
- Static Partitioning versus Dynamic Partitioning
- Partitioning makes queries go fast.
  - Partitioning works particularly well when querying with the “virtual column”
  - If queries use various columns, it may be hard to decide which columns should we partition by

# Bucketing

- Used to strike a balance between large files within partition
- Breaks up data vertically by hashed key sets
- When bucketing, you specify the number of buckets
- Works particularly well when a lot of queries contain joins
  - Especially when the two data sets are bucketed on the join key

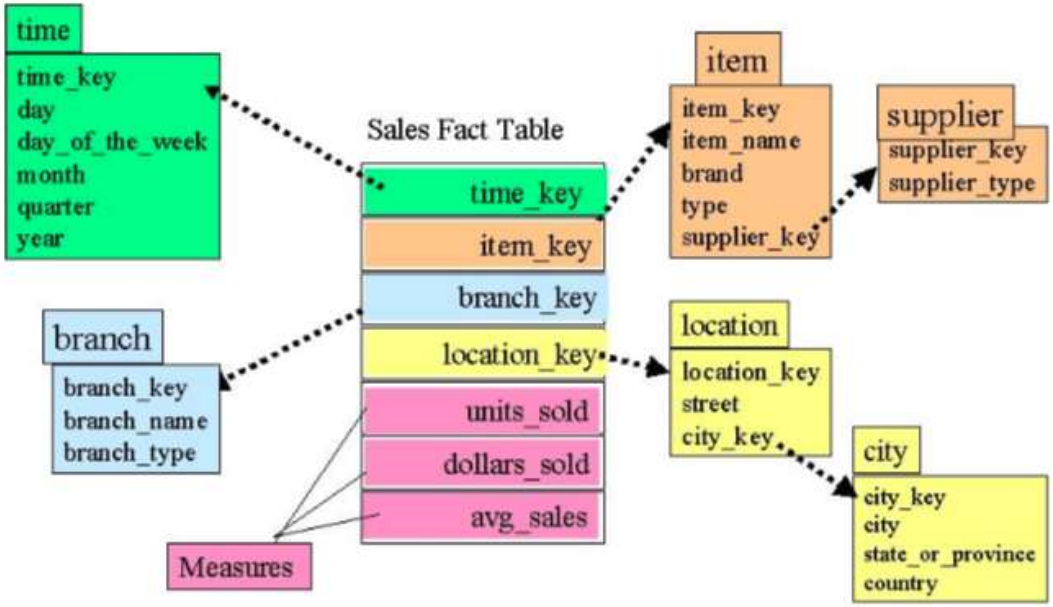
# Comparison

| Query | Text | Partition | Bucketed |
|---|---|---|---|
| select d.name, count(1), d.first_name, d.last_name from (select d.dept_no, d.dept_name as name, m.first_name as first_name, m.last_name as last_name from departments d join dept_manager dm on dm.dept_no = d.dept_no join employees m on dm.emp_no = m.emp_no where dm.to_date='9999-01-01') d join dept_emp_buck de on de.dept_no = d.dept_no join emp_buck e on de.emp_no = e.emp_no group by d.name, d.first_name, d.last_name; | 59.536 | 59.652 | 55.196 |

# Join Performance

## Map Side Joins

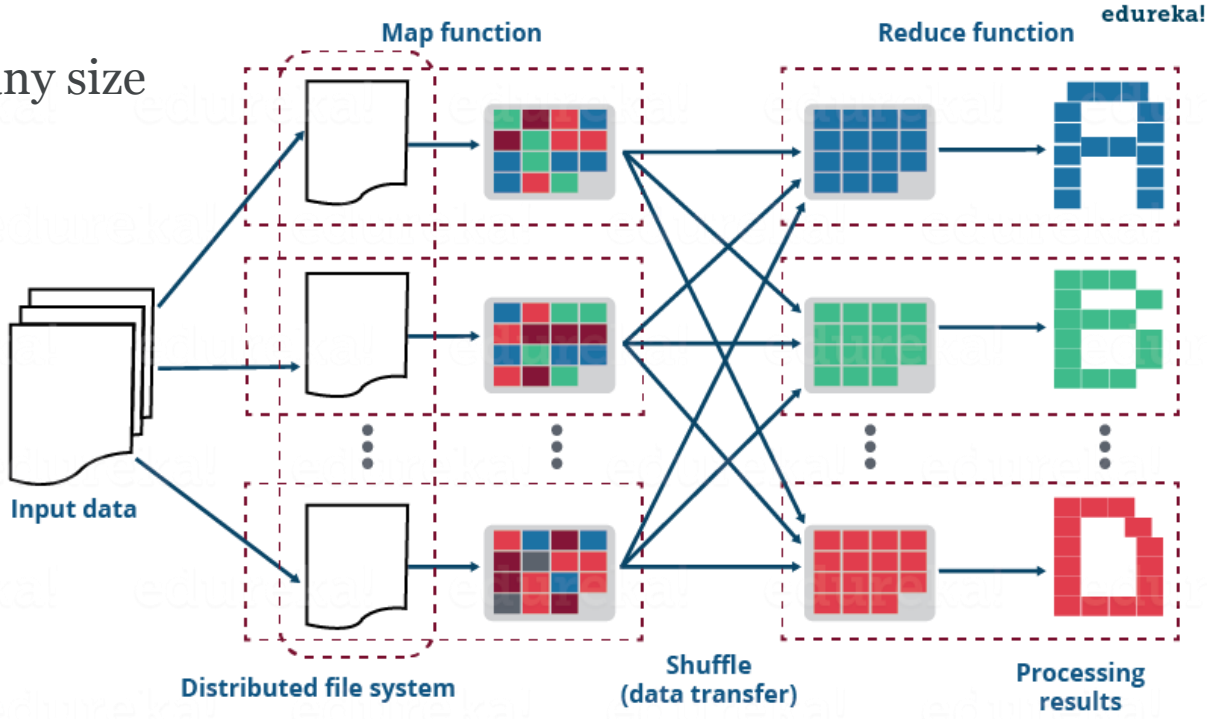
- Star schemas (e.g. dimension tables) Good when table is small enough to fit in RAM



# Reduce Side Joins

Default Hive Join

Works with data of any size

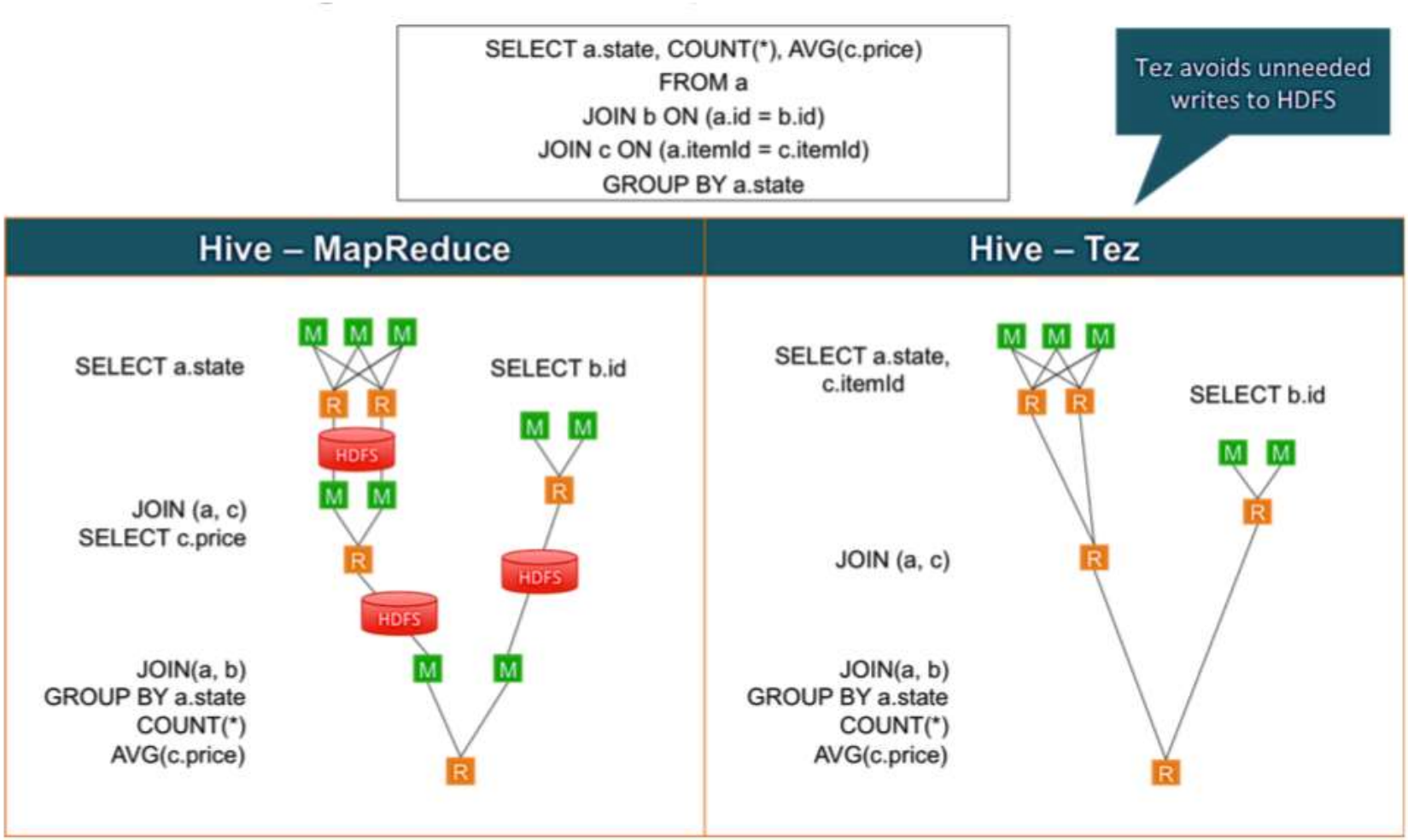


# Comparison

| Query | Map-Side | Reduce |
|---|---|---|
| select /*+ MAPJOIN(d) */ d.name, count(1), d.first_name, d.last_name from (select d.dept_no, d.dept_name as name, m.first_name as first_name, m.last_name as last_name from departments d join dept_manager dm on dm.dept_no = d.dept_no join employees m on dm.emp_no = m.emp_no where dm.to_date='9999-01-01') d join dept_emp_buck de on de.dept_no = d.dept_no join emp_buck e on de.emp_no = e.emp_no group by d.name, d.first_name, d.last_name; | 58.227 | 59.652 |

# Considerations for SQL Performance

Tez



# CBO – Cost Based Optimization

- Hive uses a Cost-Based Optimizer to optimize the cost of running a query.
- Calcite applies optimizations like query rewrite, join reordering, join elimination, and deriving implied predicates.
- Calcite will prune away inefficient plans in order to produce and select the cheapest query plans.

- Needs to be enabled:

Set hive.cbo.enable=true;

Set hive.stats.autogather=true;

## CBO Process Overview

1. Parse and validate query
2. Generate possible execution plans
3. For each logically equivalent plan, assign a cost
4. Select the plan with the lowest cost

## Optimization Factors

- Join optimization
- Table size

# LLAP

- Consists of a long-lived daemon and a tightly integrated DAG framework.
- Handles
  - Pre-fetching
  - Some Query Processing
  - Fine-grained column-level Access Control

# Daugherty Overview

- Combining **world-class capabilities** with a **local practice model**
- Long-term **consultant employees** with deep business acumen & leadership abilities
- Providing **more experienced consultants** & **leading methods/techniques/tools to:**
  - Accelerate **results** & productivity
  - Provide greater team **continuity**
  - More **sustainable/cost effective** price point.

**COLLABORATIVE**
Co-staffed teams, project Services, resource pools, collaborative managed services

**FLEXIBLE**
Flexible engagement model

**PRAGMATIC**
Pragmatic, co-staffed approach well suited to building internal competency while getting key project initiates completed

**ALTERNATIVE**
Strong Alternative to the Global Consultancies

## 9 BUSINESS UNITS

ATLANTA
CHICAGO
DALLAS
DENVER
MINNEAPOLIS
NEW YORK
SAINT LOUIS (HQ)
DEVELOPMENT CENTER
SUPPORT & HARDWARE CENTER

## BY THE NUMBERS

**1000** Over 1000 employees from Management Consultants to Developers

**88%** 88% of our clients are long-term, repeat/referral relationships of 10+ years

**75** Engagements with over 75 Fortune 500 industry leaders over the past five years

**31** Demonstrated 31 year track record of delivering mission critical initiatives enabled by emerging technologies

# Data & Analytics - What we bring to the table

**DATA & ANALYTICS**

- **Data & Analytics Strategy & Roadmap**
- **Building Analytic Solutions**
- **Analytics Competency Development**
- **Big Data / Next Gen Architecture**
- **Business Analytics and Insights**

## Methods / Tools / Techniques

- ***12 Domain EIM Blueprint/Roadmap*** framework that manages technical complexity, accelerates initiatives and focuses on delivering greatest business analytics impact quickly.
- Highly accurate ***BI Dimensional estimator*** that provides predictability in investments and time to market.
- **Analytic Strategy** framework that aligns people, process and technology components to deliver business value
- **Analytic Governance** reference model that mitigates risk and provide guardrails for self-service adoption
- **Business value models** to calculate the value and ROI of investments in Data & Analytics initiatives
- **Reference architecture** for a modern data & analytic platform
- **Dashboard Design** best practices that transform complex business KPIs in a rich immersive design
- Bi-Modal Data as a Service **Operating Model** that integrates Agile development with a Service oriented organization design

- Over ***40%*** of Daugherty's 1,000 consultants are focused on Information Management Solutions.
- Bringing the latest thought leadership in ***Next Generation, Unified Architectures*** that integrate structured, unstructured data ("Big Data") and applied advanced analytics into cohesive solutions.
- Strong **capabilities across both *existing and emerging technologies*** while maintaining a ***technology neutral*** approach.
- Leveraging the latest ***visual design*** concepts to deliver interactive and user friendly applications that drive adoption and satisfaction with solutions.
- Leader in the effective application of **Agile** techniques applied to Data Engineering development and business analytics. Full Data life cycle methods & techniques from business definition through development and on-going support
- ***Building and supporting mission-critical platforms*** for many Fortune 500 companies in multi-year, using a flexible support model including Collaborative Managed Services models.