The Rise of Big Data

Ajay Deshpande, CTO.
Rakya Technologies Pvt Ltd
www.rakya.com

Confidential Copyright 2016

INTRODUCTIONS

DAJAY0@YAHOO.COM

Getting Started…

Does a Text Book Have any Indexes?
If yes how many?

What is an Index in the context of Data Storage / Access?

How to build an Inverted Index for a document…

When would you use which one?

Confidential Copyright 2016

Basics of Data Management

Rakya

- Computers invented to store and retrieve data efficiently
- Traditional Applications: Banking, Retail, Reservations…
  - Use Traditional database systems
  - Tuned for many small simple operations
  - Online Transaction Processing (OLTP)
- Relational Database Is King
  - Transactions, ACID properties, Durability of Data
  - Entities, Relations, Tables with Rows and Columns
  - Relational Operators: Select, Project, Union, Join, etc
  - A solid foundation to manage complex data

Transactional & Structured

Confidential Copyright 2016

Basics of Data Analytics…

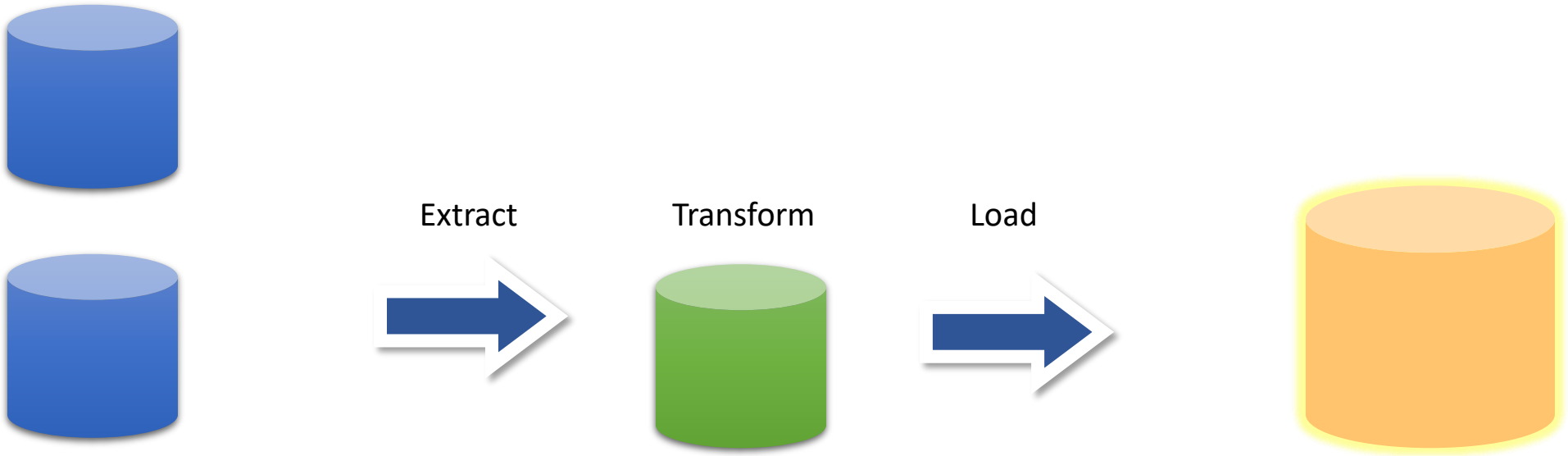
- Now that I have the data, can I understand it better?
  - Analytics: science of examining data to make decisions
- Address Questions Like
  - Top ten products sold in the last 5 years?
  - Compare the monthly totals for the last 10 years?
  - What other items are bought with Toothpaste?
- OLTP fails miserably here – why?
- Solution: Online Analytical Processing (OLAP)
  - Decision Making using the Data Warehouse

Confidential Copyright 2016

The Data Warehouse

OLTP
Daily Business
Transactional

OLAP
Reporting
Decision Making



Source Systems

Staging Area

Data Warehouse

Confidential Copyright 2016

Data Today is Ever Growing

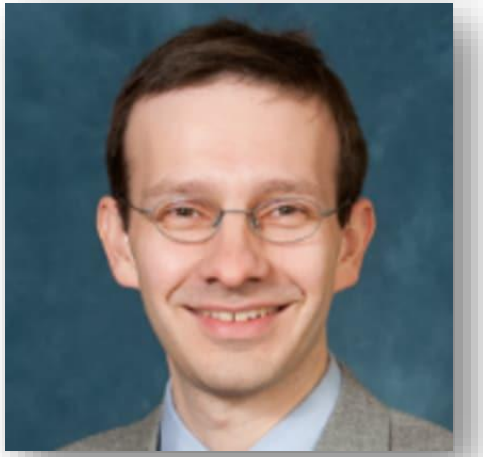
- Increase in these dimensions
  - **Velocity**: Data generated by devices / sensors
  - **Volume**: Updates on Facebook
  - **Variety**: Email, Images on Instagram
- Internet Problem: Rise of Unstructured Data
- Relational world can no longer serve
- Analysis becomes more sophisticated
  - How many people like the Tata Nano?
  - How many people will buy it next year?

Data Becomes BigData

Confidential Copyright 2016

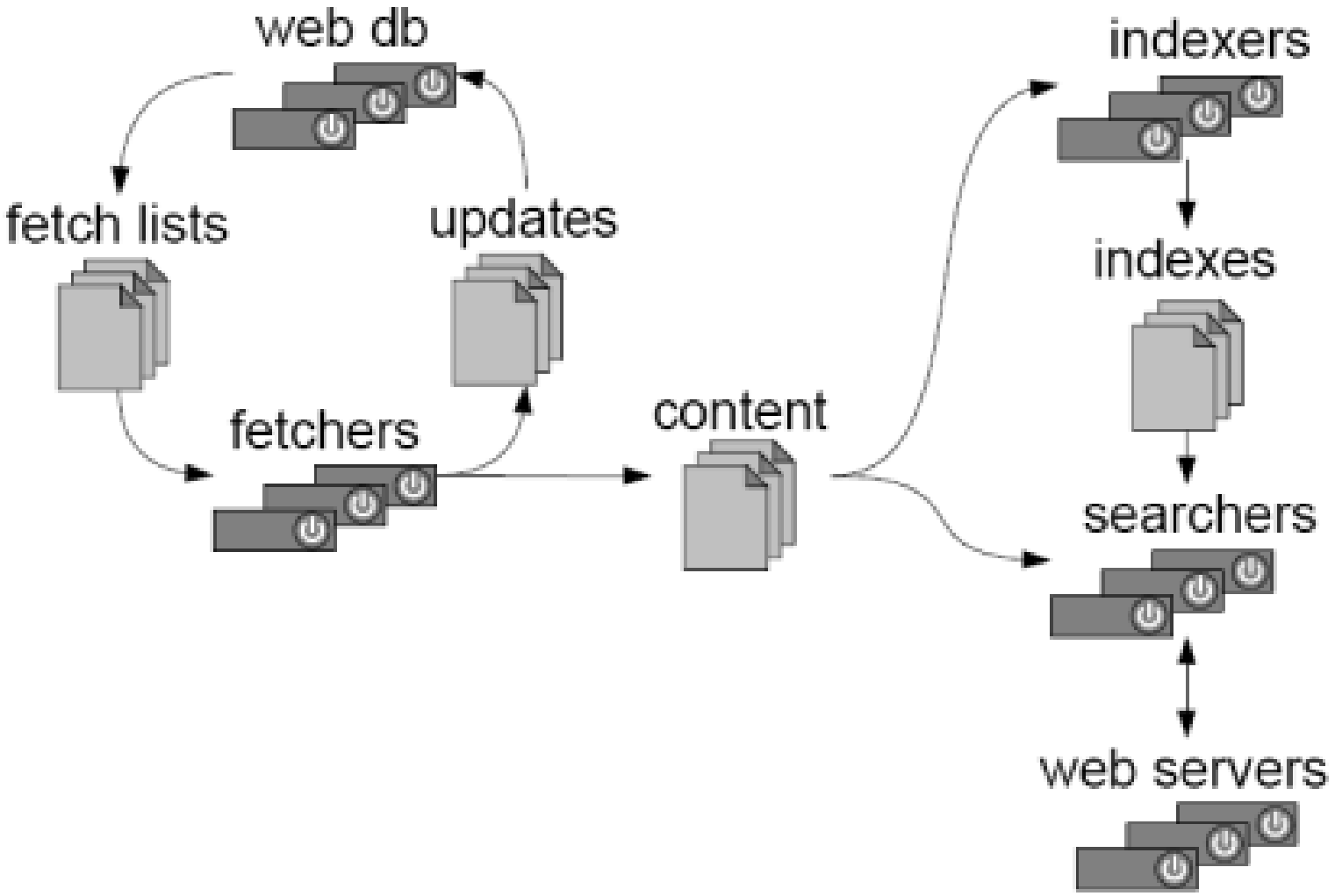
The Story Behind Hadoop

- Doug Cutting was looking to index large blocks of text – invented Lucene
- Wanted to do the same with web pages with Cafarella – Nutch was born
  - A software that could “crawl” web pages and index them
  - One machine was not sufficient. Coded to run on four nodes



Confidential Copyright 2016

The Nutch Architecture



Imagine doing this across millions of websites !

Confidential Copyright 2016

The Story Behind Hadoop…

- Became a problem trying to keep the system running
  - One or the other component kept failing
- They needed a Distributed System
  - Schema-less, Durable, Component Failure Tolerant Storage
  - And can Automatically Rebalance when nodes failed
- Google File System paper comes out – inspired NDFS

Confidential Copyright 2016

The Story Behind Hadoop…

- NDFS handled the operational issues
  - How do you distribute computation?
- Google again answered – the MapReduce paradigm
  - It handled Parallelization, Distribution and Component failure
- Instead of moving data, move the program to where the data is
- Feb 2006: Cutting pulled out NDFS + MapReduce implementation
  - Hadoop was born
- Yahoo hit a similar problem; Adopt Hadoop
  - In 2007 Cutting's team had a 1000 node cluster at Yahoo

Confidential Copyright 2016

Inside Hadoop

Introduction to Apache Hadoop

- An opensource framework to run MapReduce programs
- A platform to process large amounts of data continuously
- Execute long running computations
- Do all of this as inexpensively as possible
- Long running => Failures are inevitable => Work should not be lost

Confidential Copyright 2016

Assumptions / Goals of Hadoop

- Hardware will fail
- Tuned for batch processing / streaming data
  - Does not work well for interactive applications
- Works with huge data sets
  - Moving computation is cheaper
- Portability across platforms
- Does not allow random changes to files
  - Only Append / Truncate available
  - Enables simple concurrency control semantics
- No specialized hardware

Confidential Copyright 2016

Problem: Maximum Temperature

- You are given file(s) with data as shown
- 1000 Cities, Reading per minute for one Month
- Find Max for each City for that Month

- Solution: Process one file at a time
  - [Delhi 40] [Bengaluru 27] [Shillong 30] [Nagpur 40]
- Merge such outputs from all the files
  - Finding Max for each city after merging

Delhi | 38
Bengaluru | 27
Shillong | 30
Nagpur | 40
Delhi | 40
Bengaluru | 25

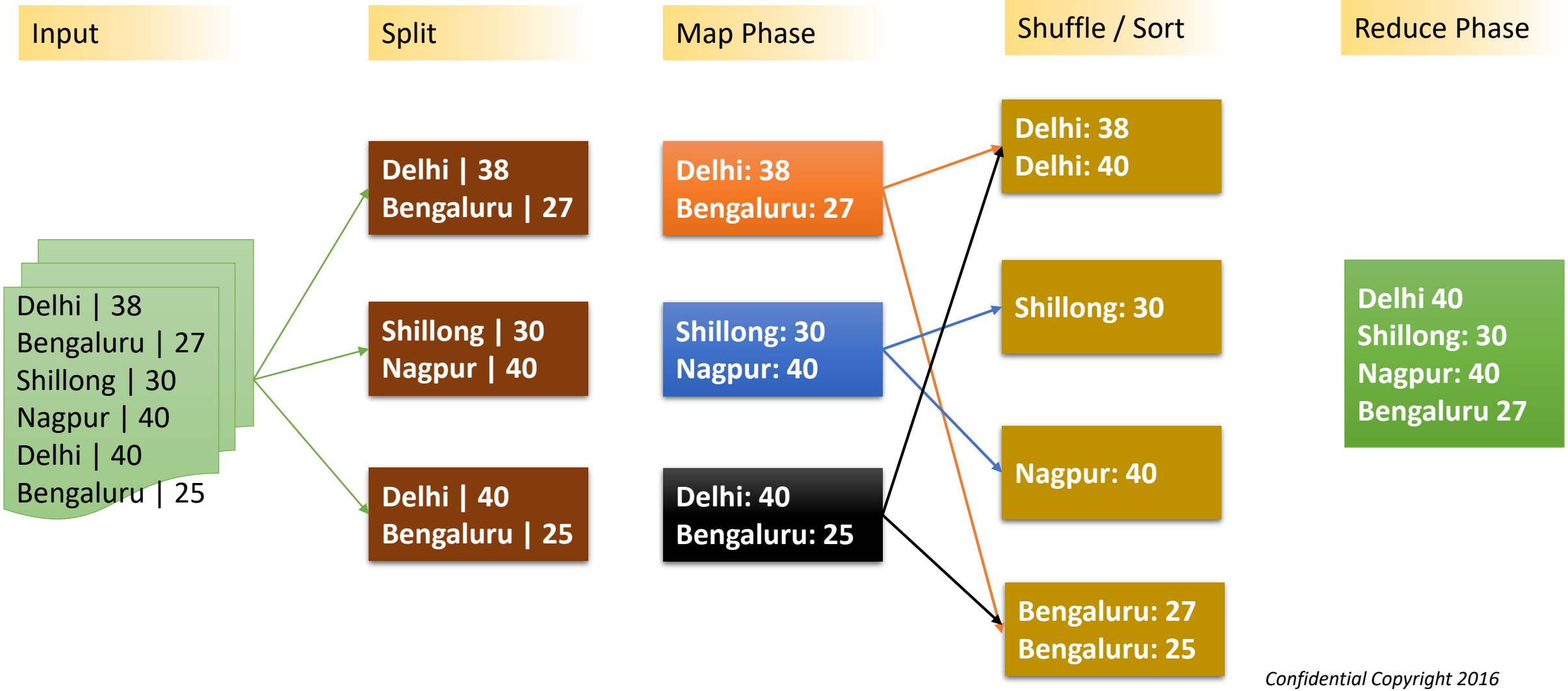
Confidential Copyright 2016

The MapReduce Paradigm

- We just used MapReduce!
- Map: List => List
  - Executes a mapping function on the input list
- Reduce: List => A single value
  - Runs a function on a list to reduce it to one value
- Divide and Conquer method
- Key to incorporating parallelism in the solution
- Increased number of nodes => Better throughput

Confidential Copyright 2016

MapReduce Applied



Confidential Copyright 2016

BigData Examples

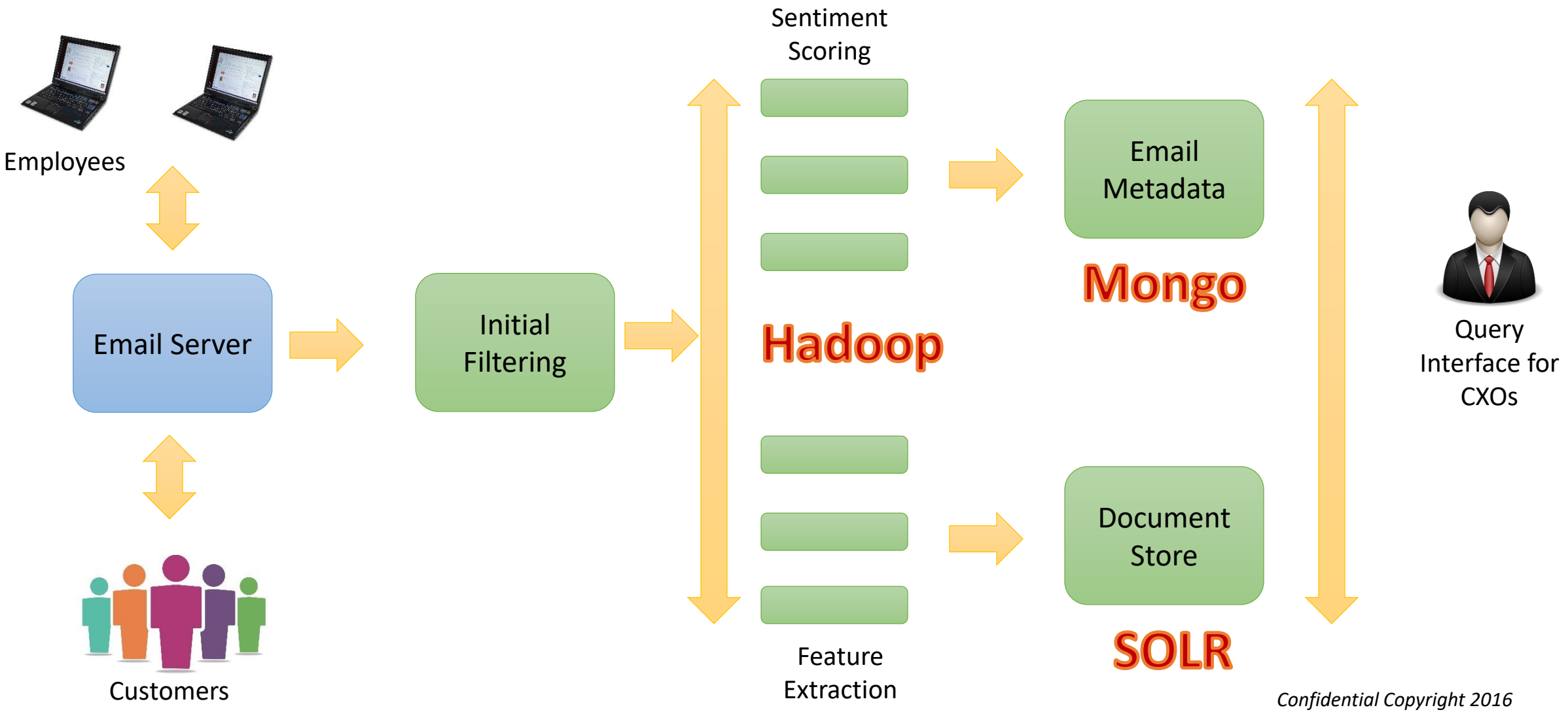
Problem: Insights From Email

- Large Company of about 10K Employees
- Primary Mode of Business Communication: Email

- CxO wants a list of Today's Unhappy Customers
- Extended to Address Other Issues
  - Making a Repository of Documents Exchanged
  - Finding Connects into customer organizations

Confidential Copyright 2016

Email Analytics for Business Insights



Confidential Copyright 2016

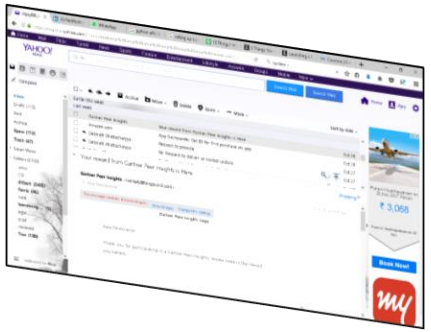
Problem: Building an Advertising Server

- Company: Provider of Digital Advertisements (Banners)
  - Works in Tandem with the Real Estate Provider
- Critical Performance Need: Return an Ad Within 200 ms

- Record Views, Click Throughs and Purchases
- Eventually Compute Payments to be made
  - Product Advertiser Pays
  - Recipients: Real Estate Owner and Advertisement Provider

Confidential Copyright 2016

Digital Advertising Server



Content

Real Estate Server

Views

Advertising Server

Cassandra

Advert Store

Clicks

Content Server

Hadoop

Purchase

MMT Server

MySQL

Admin Reports

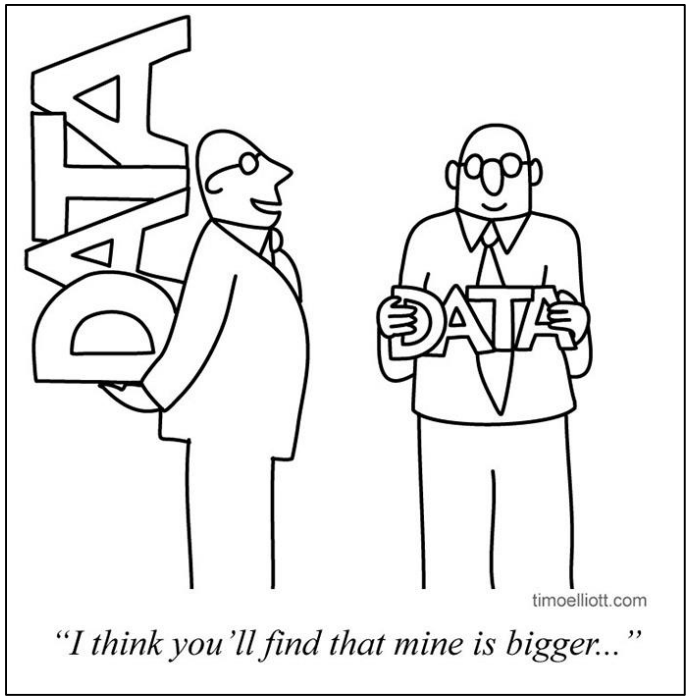
Views and Clicks Processing

Confidential Copyright 2016

Conclusions

- We are still on the Cusp of Big Data
- Data is going to only get bigger
- Best Learnt by doing hands on projects
  - Analysing Tweets, Web Logs, Emails (more ideas at **www.kaggle.com**)
- Applying Big Data to traditional domains is going to be key

Confidential Copyright 2016



THANK YOU !!

DAJAY0@YAHOO.COM