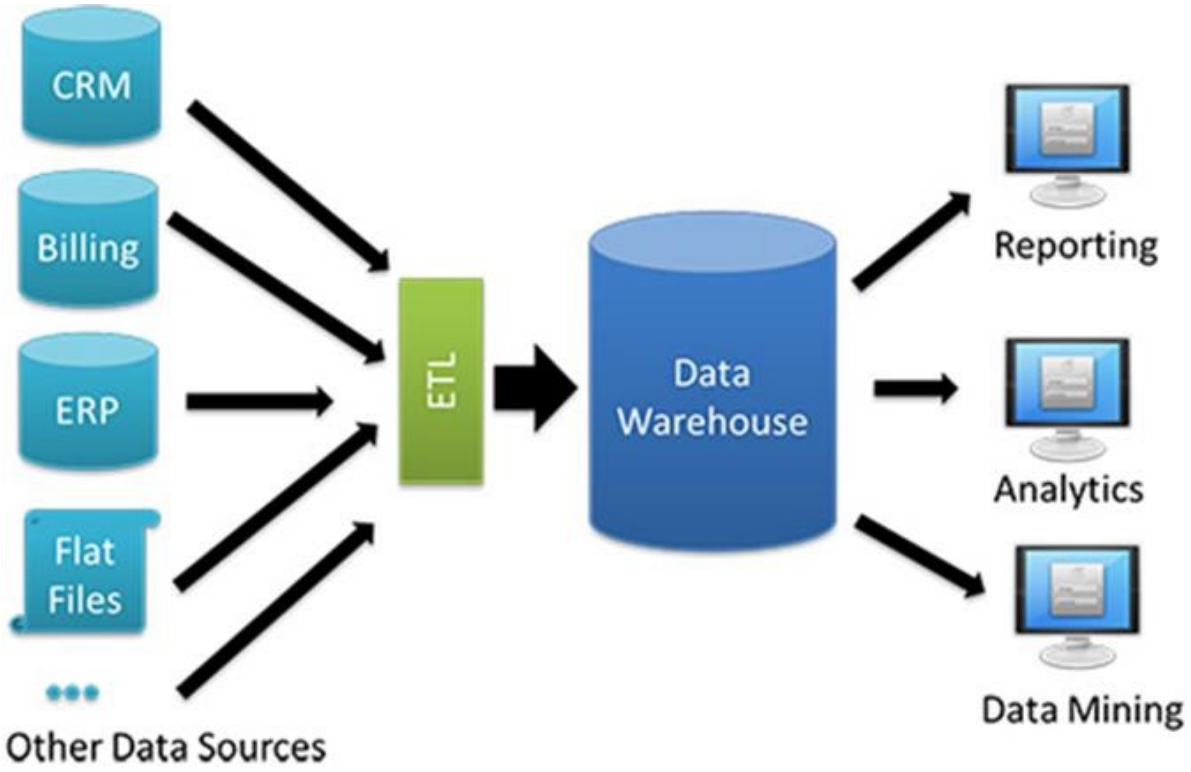
Massive Data Processing and Warehousing

Software Architecture and Scalability for Internet of Things

Dr Jonathan Kua

Data Warehouses: Overview

Gartner's simplistic view of a Data Warehouse



Data Warehouse Providers

Gartner Magic Quadrant for Data Warehouse and Data Management Solutions for Analytics - 2016

Data Warehouse Providers: Major Players

For your own knowledge, investigate the following companies data warehouse and analytics services:

These links will get you started:

https://www.monitis.com/blog/top-5-data-warehouses-on-the-market-today/

http://www.forbes.com/sites/oracle/2014/03/10/the-top-10-trends-in-data-warehousing/#19bfbe851123

Amazon Web Service: Analytics

Analytics

Business Intelligence

Amazon QuickSight

Fast Business Intelligence Service

Data Warehouse

Amazon Redshift

Fast, Simple, Cost-Effective Data Warehousing

Machine Learning

Amazon Machine Learning

Machine Learning for Developers

Streaming Data

Amazon Kinesis

Work with Real-Time Streaming Data

Elasticsearch

Amazon Elasticsearch Service

Run and Scale Elasticsearch Clusters

Hadoop

Amazon EMR

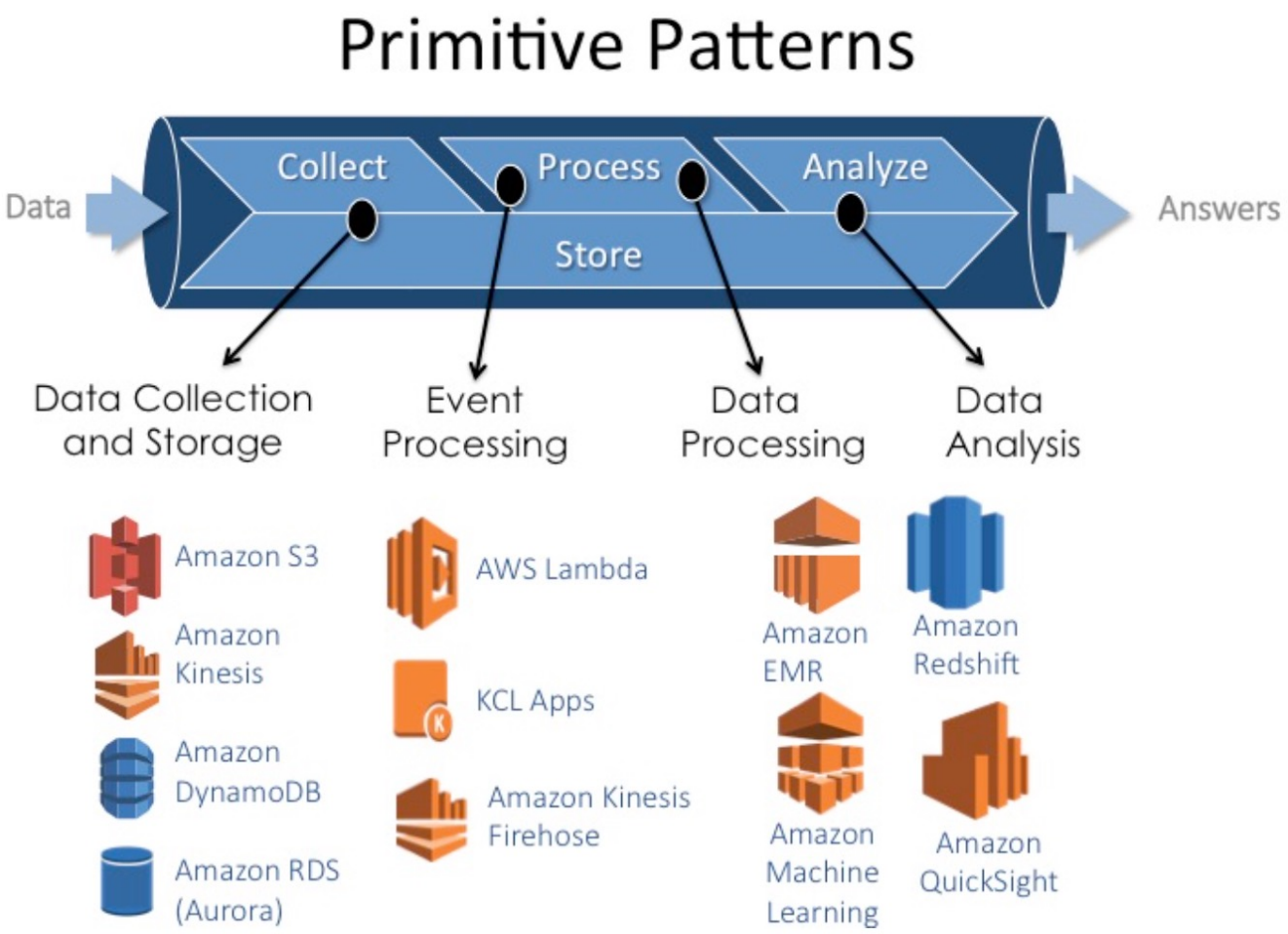
Hosted Hadoop Framework

Data Pipelines

AWS Data Pipeline

Orchestration Service for Periodic, Data-Driven Workflows

Amazon Web Service: Traditional Data Warehouses

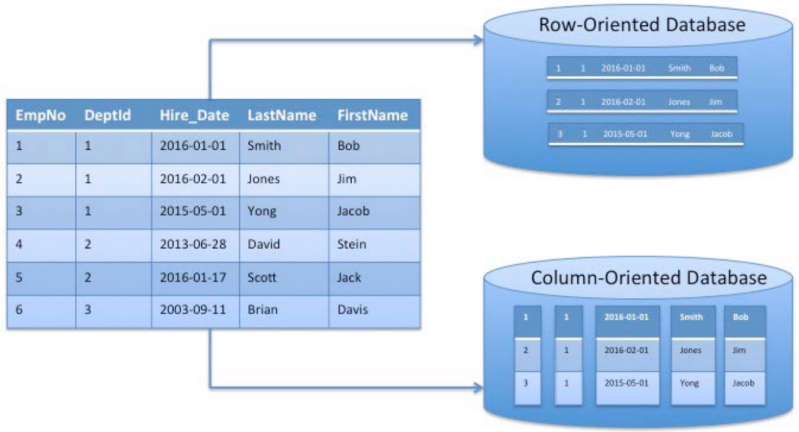


Amazon RedShift: Overview

- "Amazon Redshift is a fully managed, petabyte-scale data warehouse service in the cloud"
- Low cost
- Low Management overhead
- Launched in 2013 but has undergone rapid development
- Integrates with existing business intelligence tools
- Meant to be simple and effective

Amazon RedShift: Differences

- Column-oriented Databases
- Massively Parallel Architecture
- I/O Focused
- Scalable/Elasticity
- Interfaces for programmability
- New pipeline:



EmpNo	DeptId	Hire_Date	LastName	FirstName
1	1	2016-01-01	Smith	Bob
2	1	2016-02-01	Jones	Jim
3	1	2015-05-01	Yong	Jacob
4	2	2013-06-28	David	Stein
5	2	2016-01-17	Scott	Jack
6	3	2003-09-11	Brian	Davis

Amazon RedShift: Clusters

- An **Amazon Redshift data warehouse** is a collection of computing resources called nodes, which are organized into a group called a cluster.

- Each cluster runs an Amazon Redshift engine and contains one or more databases.

- Each cluster:

  - Leader node – receives queries from client applications
  - Compute nodes - execute queries, transmit data to leader
  - Leader node manages execution
  - (This is really just load-balancing)
  - See – CloudWatch for EC2 for something similar

Amazon RedShift: Pricing

Pay-as-you go approach (same as AWS)

"Start small for $0.25 per hour with no commitments and scale to petabytes for $1,000 per terabyte per year, less than a tenth the cost of traditional solutions. Customers typically see 3x compression, reducing their costs to $333 per uncompressed terabyte per year."

You can reserve nodes for a reduced fee

Lots of purchasing options:

http://docs.aws.amazon.com/redshift/latest/mgmt/working-with-clusters.html