Deep Diving in the World of Data Warehousing

BluePi · Follow
5 min read · Aug 23, 2017

Listen Share More

> “Information is the oil of the 21st century, and analytics is the combustion engine”. — Peter Sondergaard, Senior Vice President, Gartner.

Number of companies are venturing into Big Data Analytics today and industries are investing money, effort and looking to utilize the technical advancements for the business growth.

Till date RDBMS, excels have been favored by industries to fulfill their reporting, analytics needs, but as data is growing at rapid pace, there are roadblocks posed by these traditional measures like scalability, performance etc.

In this blog, we will try to deep dive into Analytics space and understand various terms prevalent with help of a real-world scenario.

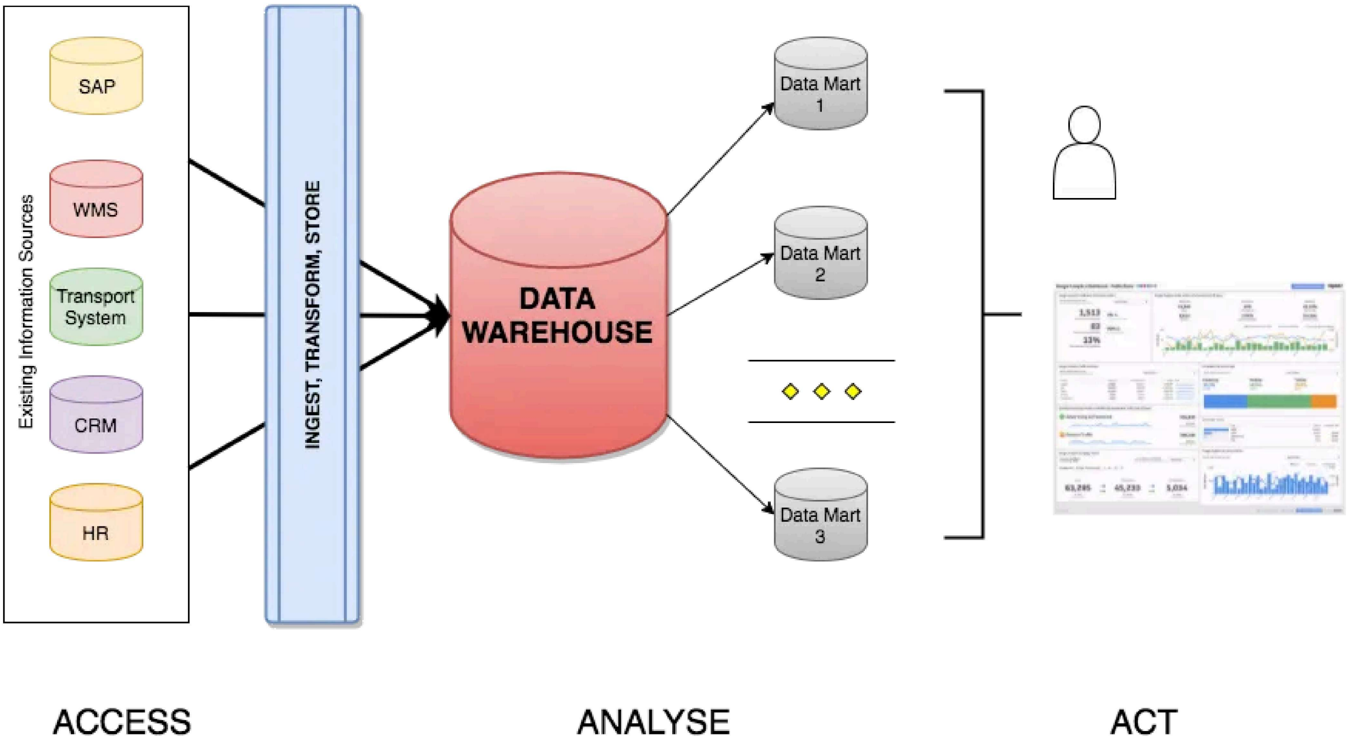
Architecture Vision

The concept of data warehousing has evolved out of the need of improving the ongoing decision making process. Over a period, organizations collect vast amounts of data but find it increasingly difficult to access and make sensible use of it. This is mainly because data is collected at different points, is in different formats, stored on many platforms. Once initial findings are made, digging deeper into data is costly, inefficient, and very time consuming.

Data warehousing helps in implementing the process of:

- Fetching data from heterogeneous data sources.
- Clean, filter, and transform the data for insights.
- Store the data in a form and manner that is easy to access, understand, and use.

The processed data is then used for query, reporting, and data analysis. As such, the access, use, technology, and performance requirements are completely different from those in a transaction-oriented operational environment.



Above diagram shows the key components of the proposed solution:

- Data Warehouse Database
- Extract, Transform, Load (ETL)

- Meta Data
- Reporting and Dashboard Tools
- Data Marts

Data Warehouse Database

A data warehouse is, by definition, a subject-oriented, integrated, time-variant collection of data.

There are number of solutions present in market like Teradata's EDW Platform, Oracle Exadata Machine, or our in-house favourite and most competent of them all Amazon Redshift.

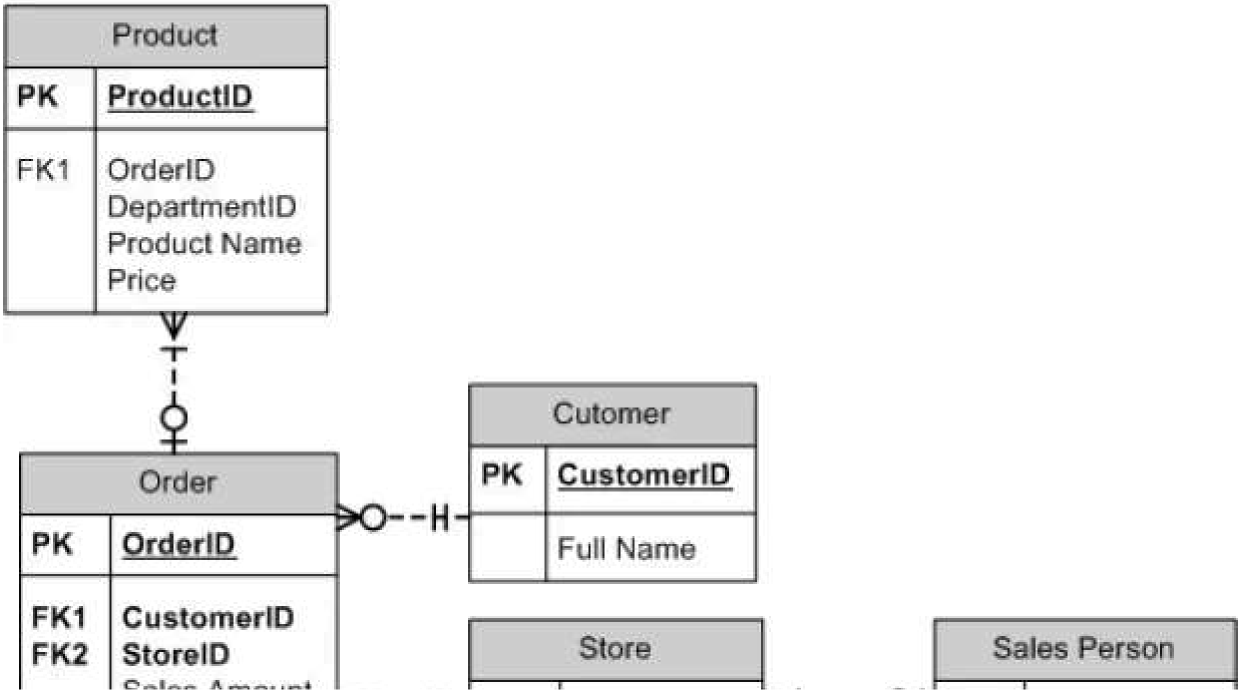
One of the most basic concepts of data warehousing is to clean, filter, transform, summarize, and aggregate the data, and then put it in a structure for easy access and analysis by those users. But, that structure must first be defined and that is the task of the data warehouse model.

There are two basic data modeling techniques

- ER Modeling
- Dimensional Modeling

ER Modelling

ER Modelling consists of three basic concepts i.e. entities, attributes and relationships.

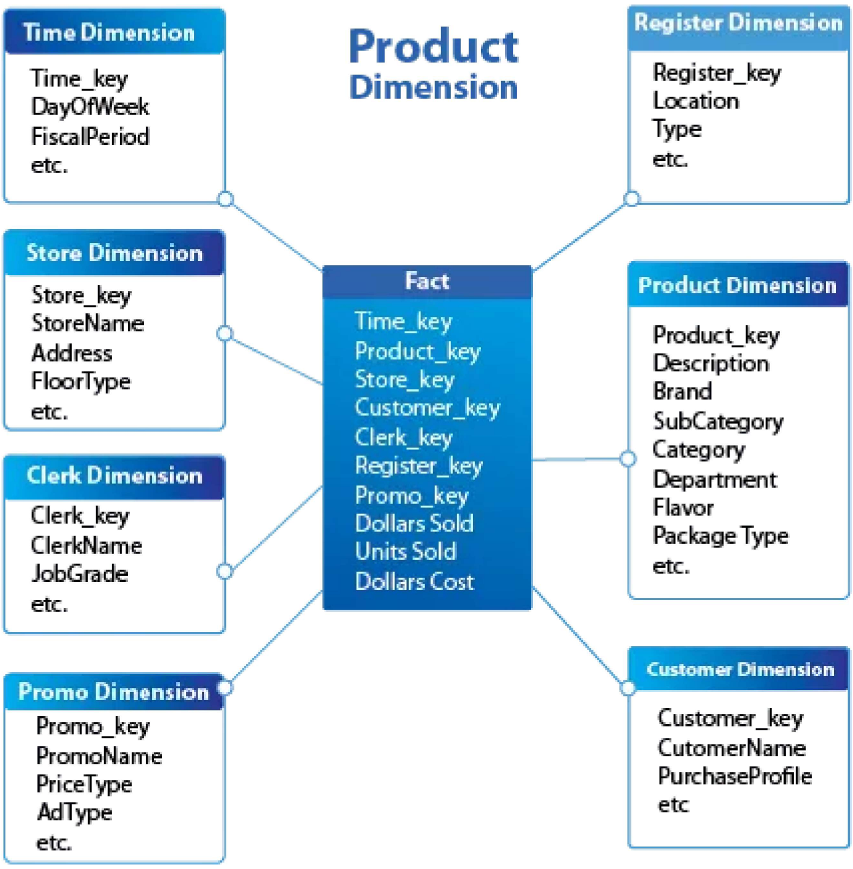


Its better when we want to keep data in normalized form and since entities are the core concept the data load and query revolves around the logical, real world objects like Product, Customer etc.

This data modelling is followed by the RDBMS data stores and are very efficient for the transaction based systems but fail to scale when comes to analytical systems.

Dimensional Modelling

Dimensional modeling is simpler, more expressive, and easier to understand. There are 3 basic concepts in dimensional modeling i.e. **facts**, **dimensions** and **measures.**



A fact is a collection of related data items, which consist of measures and context data. Each fact typically represents a business item, a business transaction, or an event that can be used in analyzing the business or business processes.

A dimension is a collection of members or units of the same type of views. Dimensions are the parameters over which we want to perform Online Analytical Processing (OLAP). For example, in a database for analyzing all sales of products, common dimensions could be:

- Time
- Location/region
- Customers
- Salesperson

A measure is a numeric attribute of a fact, representing the performance or behavior of the business relative to the dimensions. For example, measures are the sales in money, the sales volume, the quantity supplied, the supply cost, the transaction amount etc.

Dimensional modeling is primarily used to support OLAP and decision making while ER modeling is best fit for OLTP where results consist of detailed information of entities rather an aggregated view.

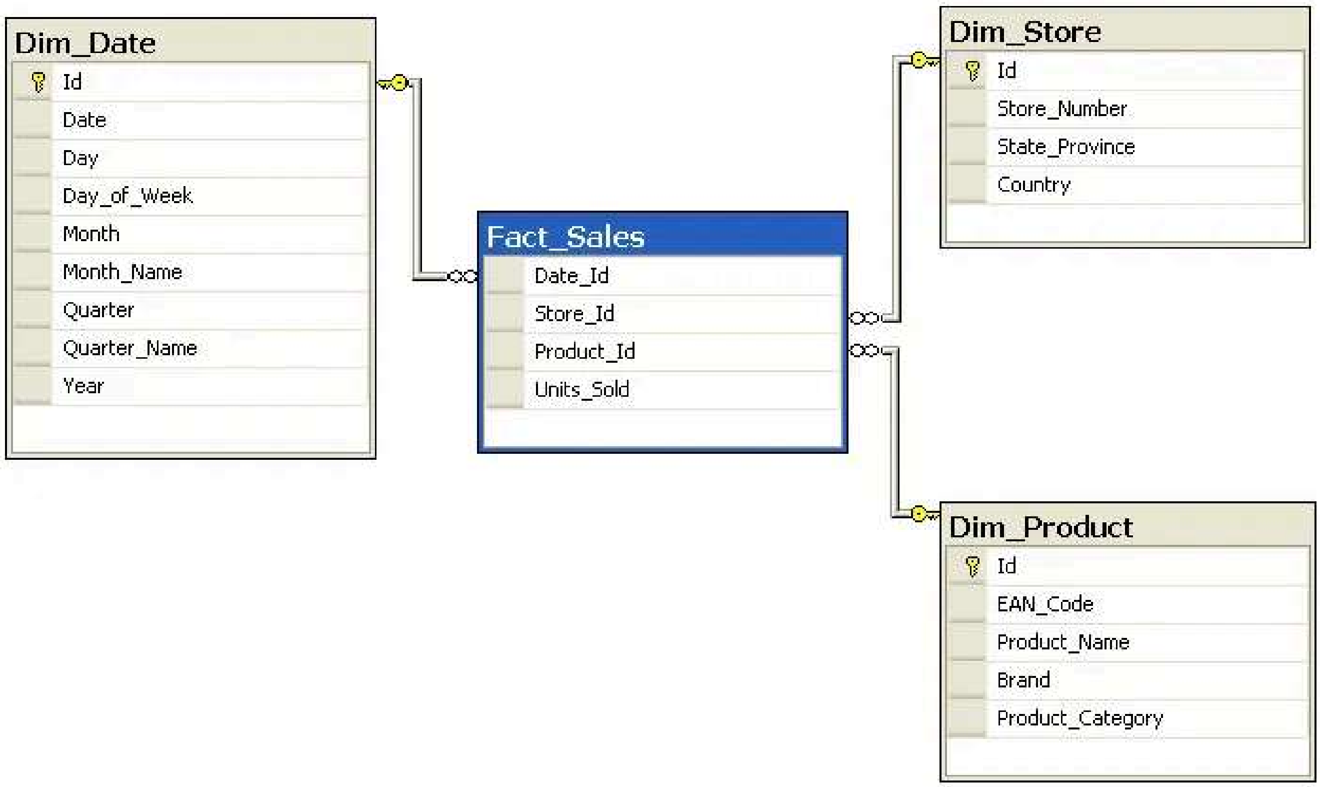
It provides four types of operations: **Drill down**, **Roll up**, **Slice** and **Dice**.

Drill down and roll up are the operations for moving the view down and up along the dimensional hierarchy levels to get more refined or bird-eye views. With drill-down capability, users can navigate to higher levels of detail. With roll-up capability, users can zoom out to see a summarized level of data.

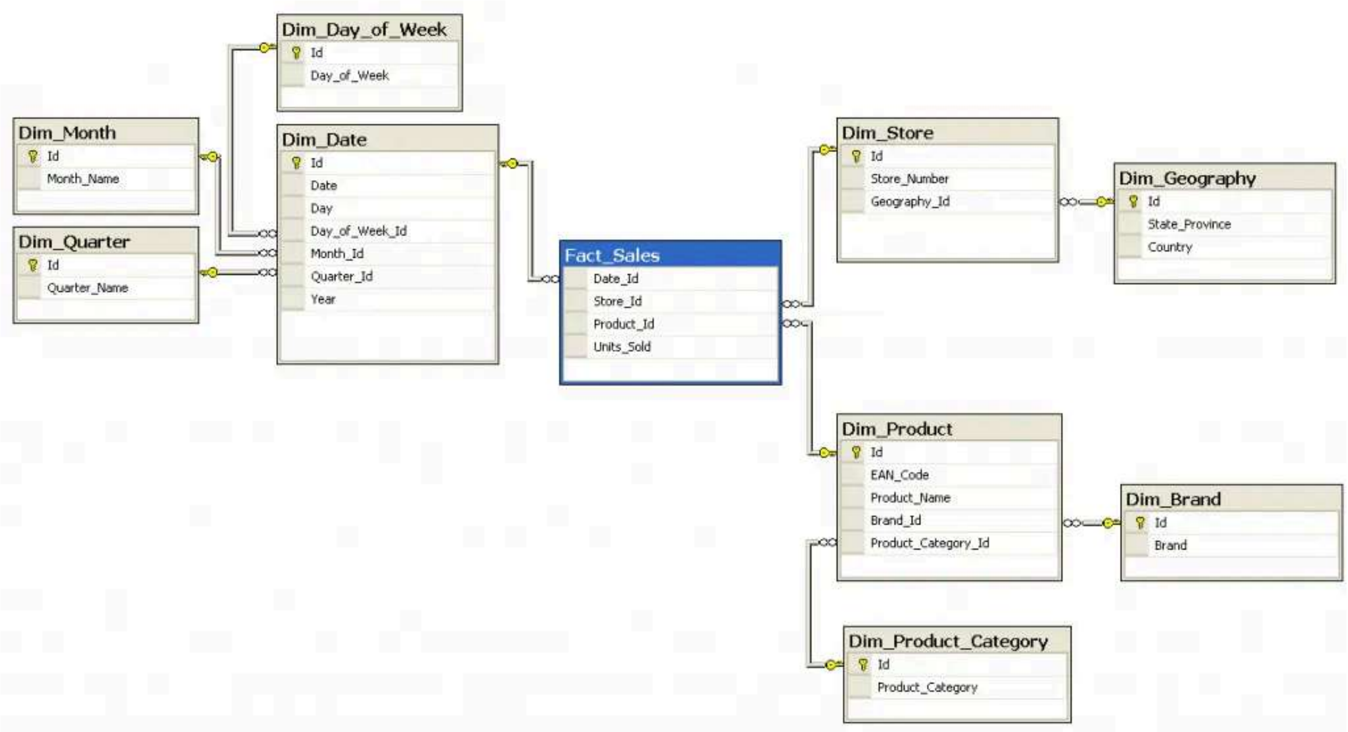
Slice and dice are the operations for browsing the data through the visualized cube. Slicing cuts through the cube so that users can focus on some specific perspectives. Dicing rotates the cube to another perspective so that users can be more specific with the data analysis.

Star and Snowflake Schema

The star model is the basic and most widely used structure for a dimensional model. It typically has one large central table (called the fact table) and a set of smaller tables (called the dimension tables) arranged around the fact table, as shown below.



The snowflake model is like star schema but results from decomposing one or more of the dimensions, which sometimes have hierarchies themselves. This leads to normalization of data but may increase query cost. Example is shown below:



Hope that with our detailed analysis and description, you have gathered some interesting and knowledgeable insights, and now these various terminologies used in a typical IT setup, imparts you more clarity and correct usage.

In our next blog of this series, we will see various other components used to build an end-to-end analytics solution.

Till then, keep innovating!

Big Data Data Warehouse Amazon Redshift Redshift Database Development

Follow

Written by BluePi

287 Followers · 346 Following

Premier #AWS Consulting Partner. #AI/ML driven solutions for #retail #QSR. #MicrosoftPartners #RetailAnalytics #CloudServices #DataScience

No responses yet

Giorgio Zoppi

What are your thoughts?

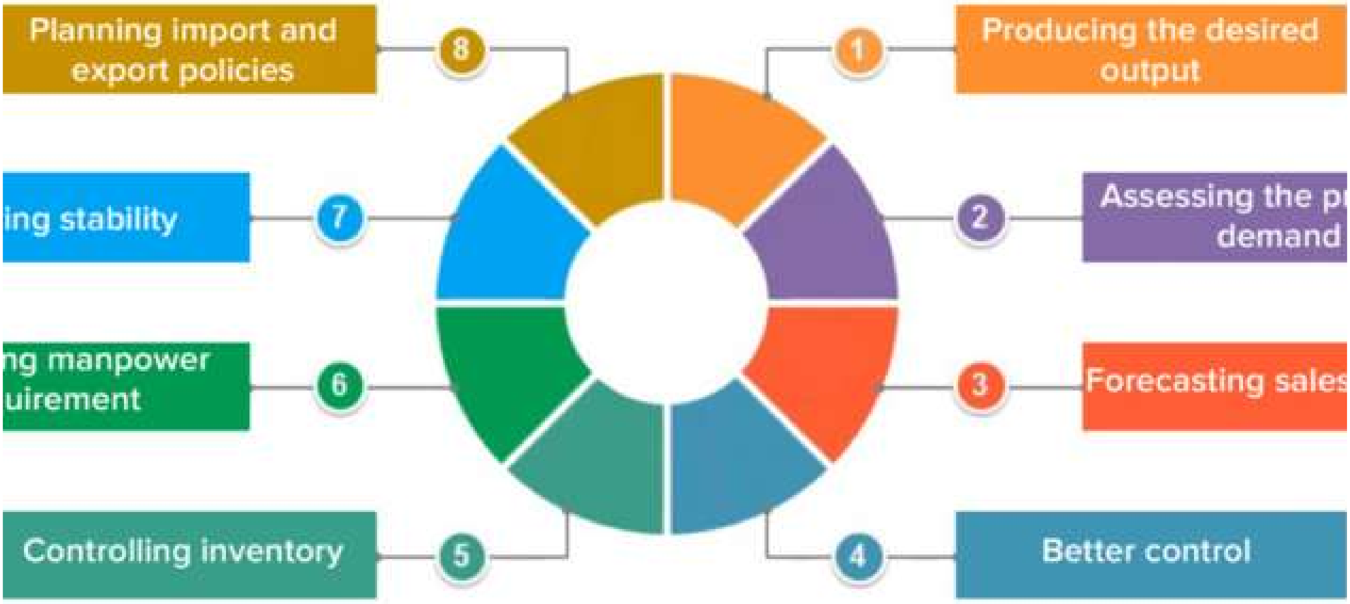
More from BluePi

BluePi

Deep Diving in the World of Data Warehousing — Continued..

In our last blog, we tried to understand terminology around data warehouse. But data warehouse is only a central storage system, to build a...

Sep 20, 2017 1



In BluePi Blog by BluePi

What are the different demand forecasting techniques?

Introduction

Nov 26, 2020 22

BluePi

Live Webinar — Ansible Automation

Ansible Automation poses big advantages over other tools as it 1) creates efficiency, 2) ensures security, and 3) provides scalability with...

May 11, 2016 1

In BluePi Blog by BluePi

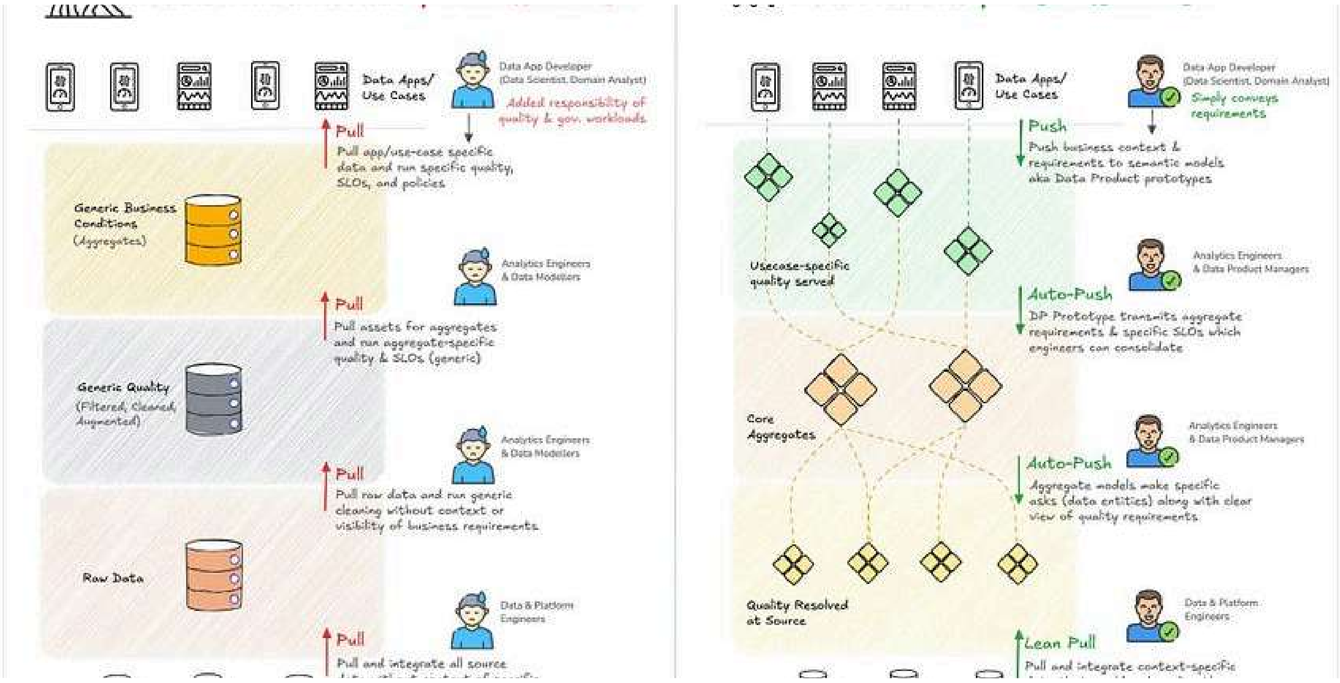
Role of third party data in a customer data platform

If there is one sector that has remained buoyant despite the global pandemic, it is online commerce. Brick and mortar players are also a...

Nov 5, 2020

See all from BluePi

Recommended from Medium

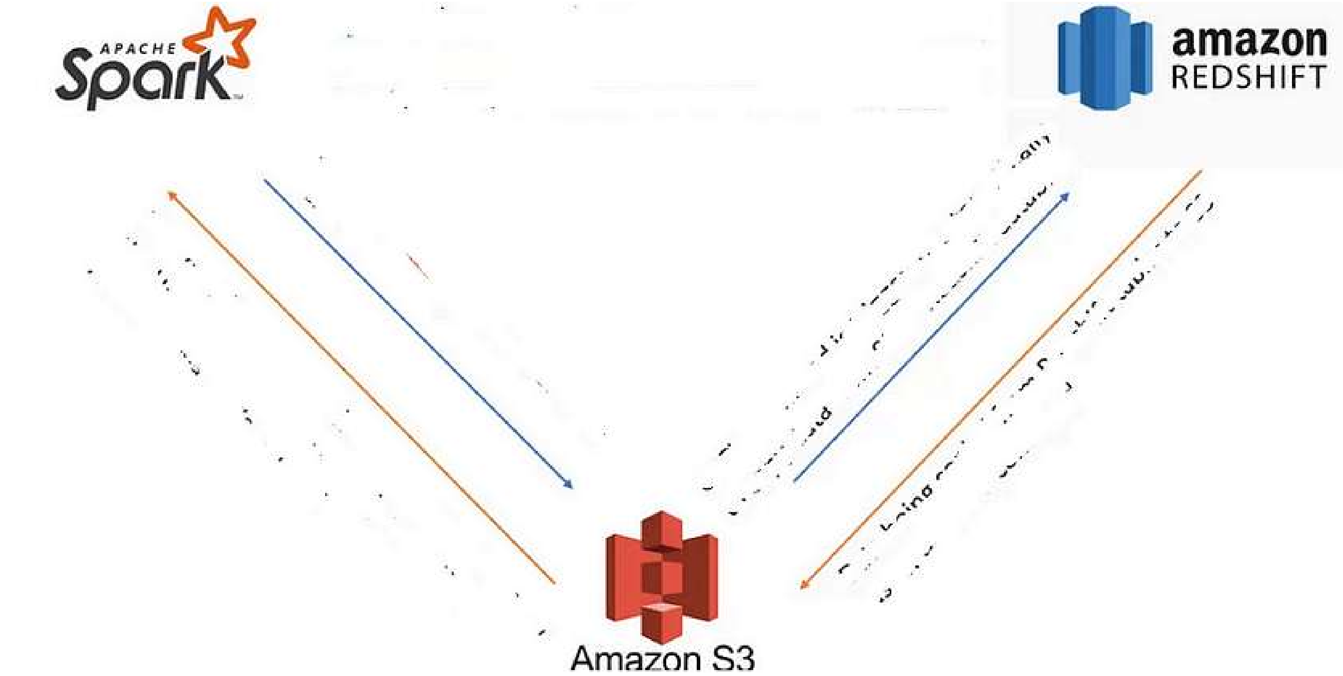

Data Apps/ Use Cases
Data App Developer (Data Scientist, Domain Analyst)
Added responsibility of quality & gov. workloads
Pull
Pull app/use-case specific data and run specific quality, SLOs, and policies
Generic Business Conditions (Aggregates)
Analytics Engineers & Data Modellers
Pull
Pull assets for aggregates and run aggregate-specific quality & SLOs (generic)
Generic Quality (Filtered, Cleaned, Augmented)
Analytics Engineers & Data Modellers
Pull
Pull raw data and run generic cleaning without context or visibility of business requirements
Raw Data
Data & Platform Engineers
Pull
Pull and integrate all source
Data Apps/ Use Cases
Data App Developer (Data Scientist, Domain Analyst)
Simply conveys requirements
Push
Push business context & requirements to semantic models aka Data Product prototypes
Usecase-specific quality served
Analytics Engineers & Data Product Managers
Auto-Push
DP Prototype transmits aggregate requirements & specific SLOs which engineers can consolidate
Core Aggregates
Analytics Engineers & Data Product Managers
Auto-Push
Aggregate models make specific asks (data entities) along with clear view of quality requirements
Quality Resolved at Source
Data & Platform Engineers
Lean Pull
Pull and integrate context-specific

Modern Data 101

Data Products: A Case Against Medallion Architecture

The Significance of Medallion, Crux of the Differences between the two 3-Tiered DataFlow Models, and a Colourful Visual Journey!

Feb 18 426 16



azeem jainulabdeen

PySpark (on S3) vs Redshift:

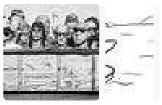
When using PySpark to process data stored in Amazon S3 instead of using Amazon Redshift, you will be working in different paradigms, and…

Sep 19, 2024

Lists

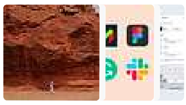
data science and AI

40 stories · 337 saves



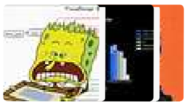
Staff picks

819 stories · 1637 saves



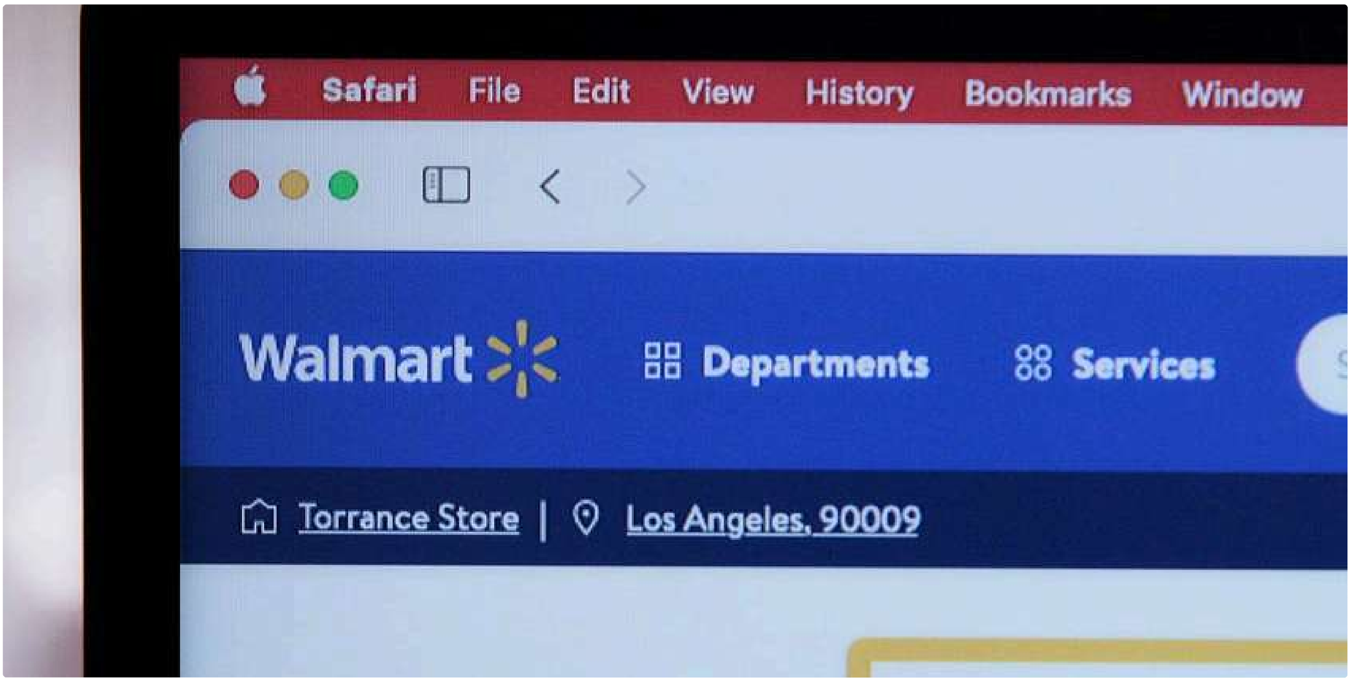
Interesting Design Topics

258 stories · 972 saves



Natural Language Processing

1958 stories · 1605 saves

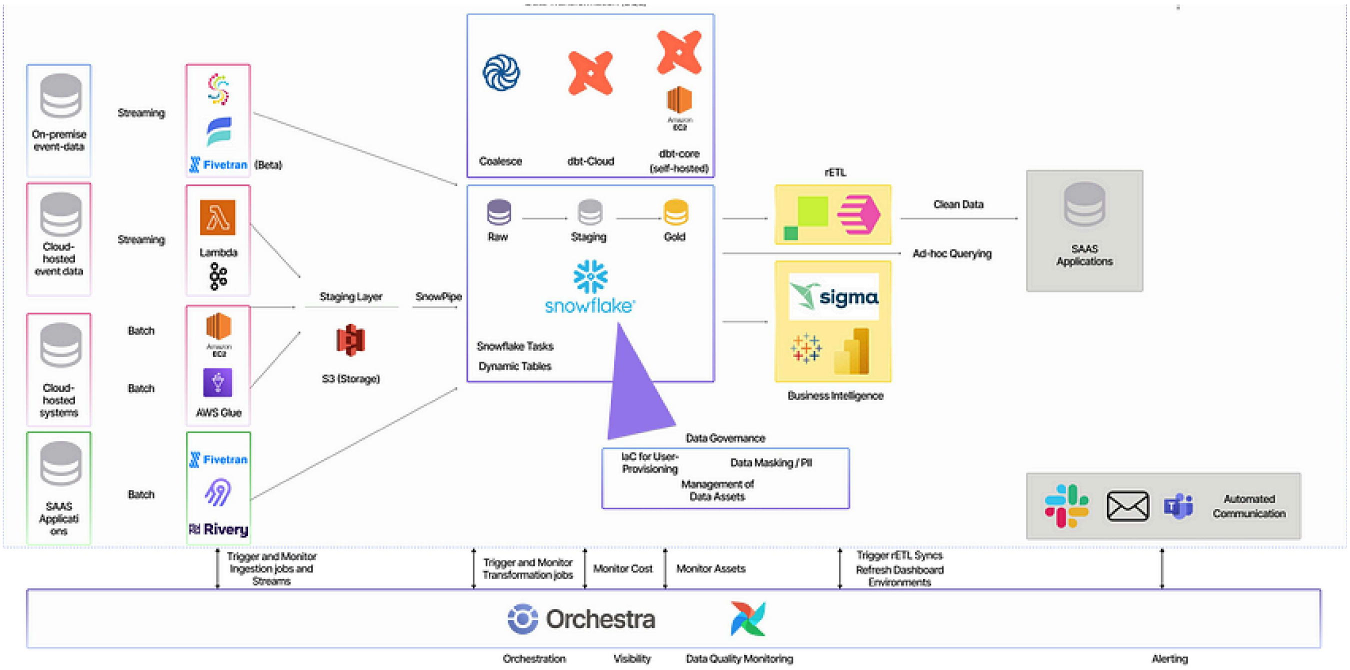


Think Data

My Interview Experience for Senior Data Engineer at Walmart

I’m here to share my Senior Data Engineer interview experience with Walmart from December 2024. If you’re preparing for this role, I hope...

6d ago 103 4



Hugo Lu

Enterprise Data Architecture 101: AWS+Snowflake Blueprints

A framework for understanding Enterprise Data Architecture on AWS in Snowflake for 2024

Sep 27, 2024 100 2

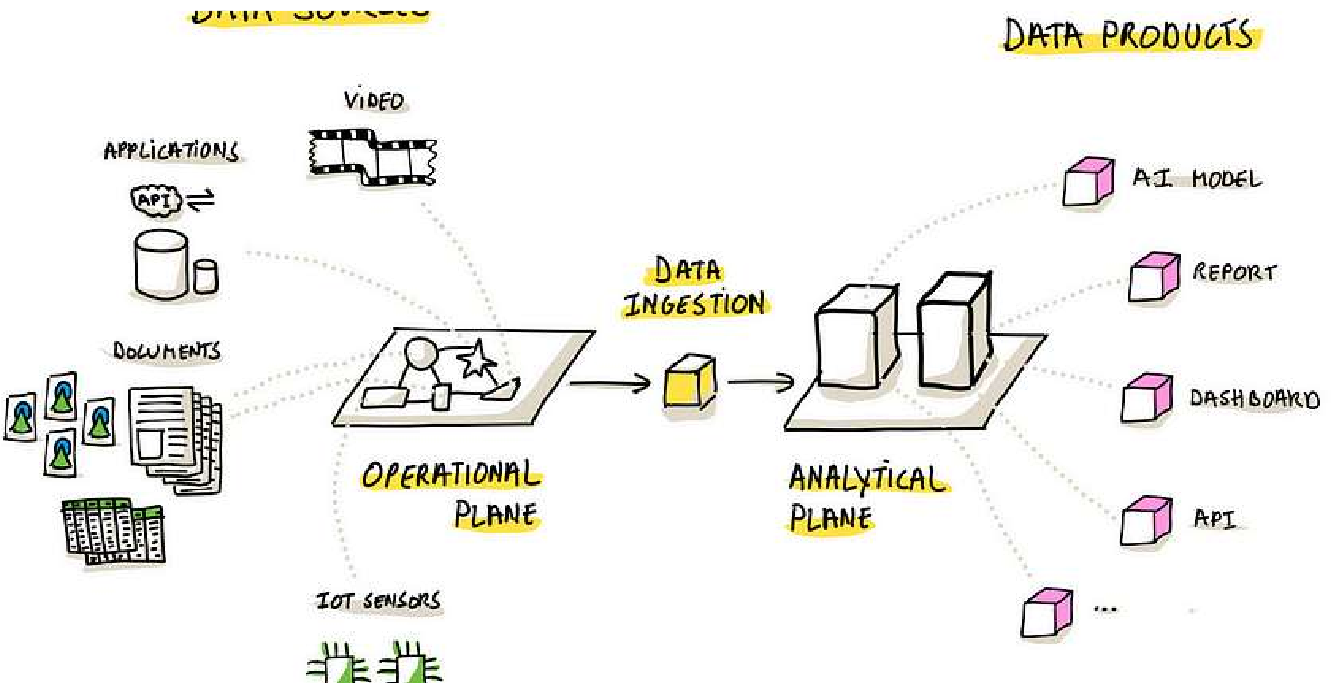
On the blog
AALAM
GETTING STARTED WITH SSIS: A BEGINNER'S GUIDE TO DATA INTEGRATION USING VISUAL STUDIO
WRITTEN BY KISHORE KUMAR CR

Aalam Info Solutions LLP

Getting Started with SSIS: A Beginner's Guide to Data Integration Using Visual Studio

SQL Server Integration Services (SSIS) is a powerful tool for data integration and using it effectively can significantly streamline user's...

Jan 17



In The Modern Scientist by janmeskens

Data Ingestion — Part 1: Architectural Patterns

Over the course of two articles, I will thoroughly explore data ingestion, a fundamental process that bridges the operational and...

Nov 27, 2023 2.7K 24

See more recommendations