Architecture Design

Analysing Google Apps Store dataset

Presented By:

Ishita Shetty

Document Version Control

Date Issued	Version	Description	Author
17th June 2022	1.0	Introduction, Architecture, Deployment	Ishita Shetty
23rd June 2022	1.1	Final Revision	Ishita Shetty

Contents

1 Introduction

1.1 What is Architecture Design Document?

Any software needs the architectural design to represents the design of software. IEEE defines architectural design as “the process of defining a collection of hardware and software components and their interfaces to establish the framework for the development of a computer system.” The software that is built for computer-based systems can exhibit one of these many architectures.

Each style will describe a system category that consists of :

- A set of components (eg: a database, computational modules) that will perform a function required by the system.
- The set of connectors will help in coordination, communication, and cooperation between the components.
- Conditions that how components can be integrated to form the system.

Semantic models that help the designer to understand the overall properties of the system.

1.2 Scope

Architecture Design Document (ADD) is an architecture design process that follows a step-by-step refinement process. The process can be used for designing data structures, required software architecture, source code and ultimately, performance algorithms. Overall, the design principles may be defined during requirement analysis and then refined during architectural design work.

2 Architecture

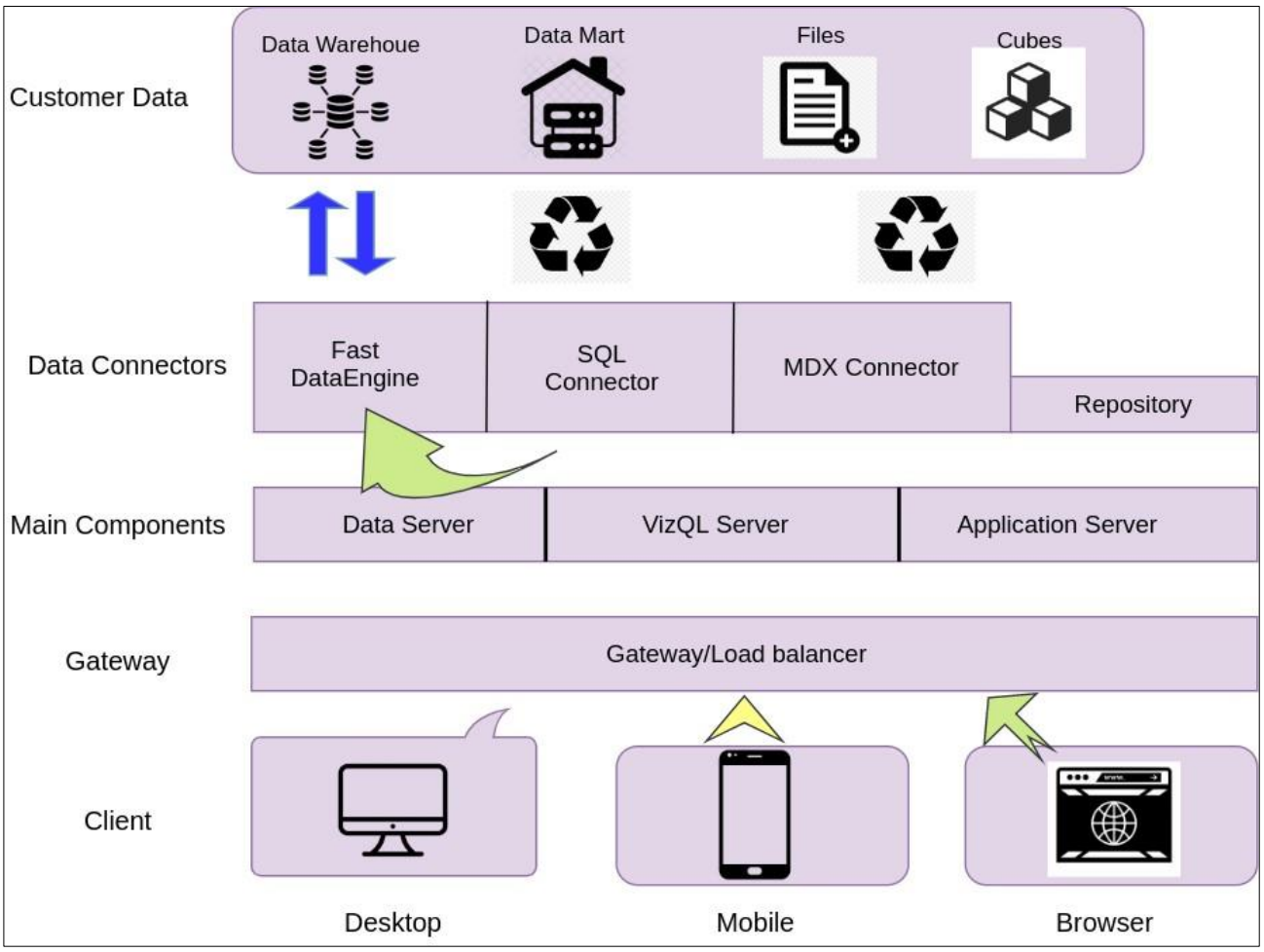


Figure 1. Working of BI

2.1 Working of BI

1. **Raw Data Collection:**

The Dataset was taken from iNeuron's Provided Project Description Document.
https://drive.google.com/drive/folders/165Pjmfb9W9PGy0rZjHEA22LW0Lt3Y-Q8

2. **Data Pre-Processing:**

Before building any model, it is crucial to perform data pre-processing to feed the correct data to the model to learn and predict. Model performance depends on the quality of data feeded to the model to train.
This Process includes:
a) Handling Null/Missing Values

b) Handling Skewed Data
c) Outliers Detection and Removal

3. Data Cleaning:

Data cleaning is the process of fixing or removing incorrect, corrupted, incorrectly formatted, duplicate, or incomplete data within a dataset.
a) Remove duplicate or irrelevant observations
b) Filter unwanted outliers
c) Renaming required attributes

4. Exploratory Data Analysis (EDA)

Exploratory Data Analysis refers to the critical process of performing initial investigations on data to discover patterns, spot anomalies, test hypothesis and to check assumptions with the help of summary statistics and graphical representations.

5. Reporting

Reporting is a most important and underrated skill of a data analytics field.
Because being a Data Analyst you should be good in easy and self explanatory report because your model will be used by many stakeholders
who are not from technical background.
a) High Level Design Document (HLD)
b) Low Level Design Document (LLD)
c) Architecture
d) Wireframe
e) Detailed Project Report
f) Power Point Presentation

6. Modelling

Data Modelling is the process of analysing the data objects and their relationship to the other objects. It is used to analyse the data requirements that are required for the business processes. The data models are created for the data to be stored in a database. The Data Model's main focus is on what data is needed and how we have to organize data rather than what operations we have to perform.

7. Deployment

We created a Dashboard on Tableau

2.2 Tableau Server Architecture

Tableau has a highly scalable, n-tier client-server architecture that serves mobile clients, web clients and desktop-installed software. Tableau Server architecture supports fast and flexible deployments.

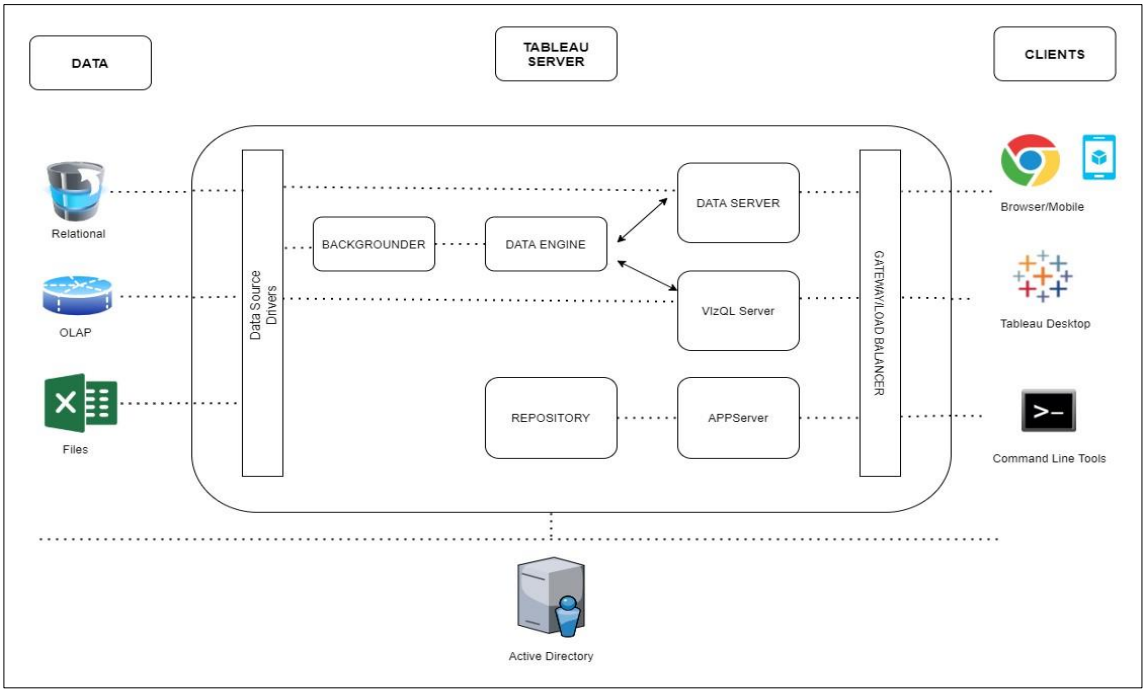


Figure 2. Architecture of Tableau Server

Tableau Server is internally managed by the multiple server processes.

1. **Gateway/Load Balancer:**

It acts as an Entry gate to the Tableau Server and also balances the load to the Server if multiple Processes are configured.

2. **Application Server:**

Application Server processes (wgserver.exe) handle browsing and permissions for the Tableau Server web and mobile interfaces. When a user opens a view in a client device, that user starts a session on Tableau Server. This means that an Application Server thread starts and checks the permissions for that user and that view.

3. Repository:

Tableau Server Repository is a PostgreSQL database that stores server data. This data includes information about Tableau Server users, groups and group assignments, permissions, projects, data sources, and extract metadata and refresh information.

4. Data Engine:

It stores data extracts and answers queries

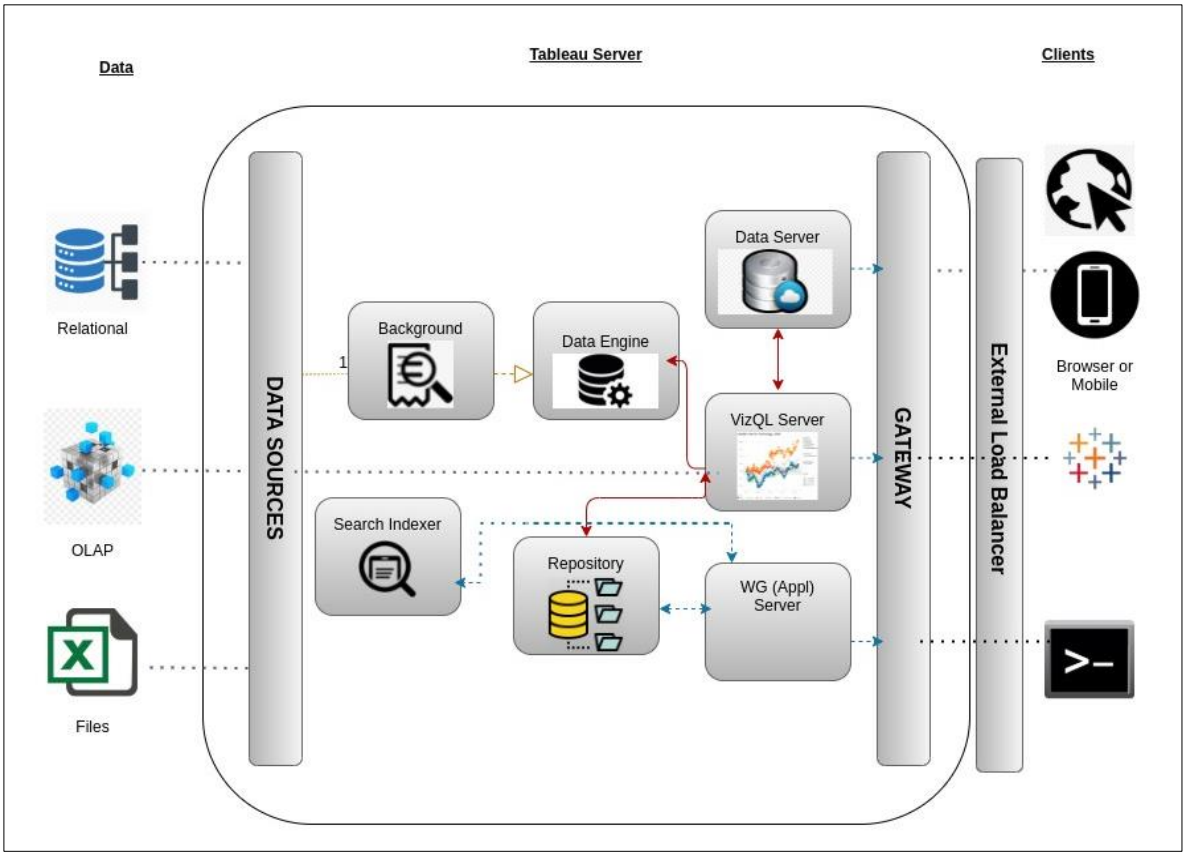
5. Backgrounder:

The backgrounder Executes server tasks which includes refreshes scheduled extracts, tasks initiated from tabcmd and manages other background tasks.

6. Data Server:

Data Server Manages connections to Tableau Server data sources It also maintains metadata from Tableau Desktop, such as calculations, definitions, and groups.

7. Tableau Communication Flow:



3 Deployment Description

Prioritizing data and analytics couldn't come at a better time. A company, no matter what size, is already collecting data and most likely analysing just a portion of it to solve business problems, gain competitive advantages, and drive enterprise transformation. With the explosive growth of enterprise data, database technologies, and the high demand for analytical skills, today's most effective IT organizations have shifted their focus to enabling self-service by deploying and operating Tableau at scale, as well as organizing, orchestrating, and unifying disparate sources of data for business users and experts alike to author and consume content.

3.1 Deployment options in Tableau

Tableau's analytics platform offers three different deployment options depending on your environment and needs. The below graphic shows each option at a glance:

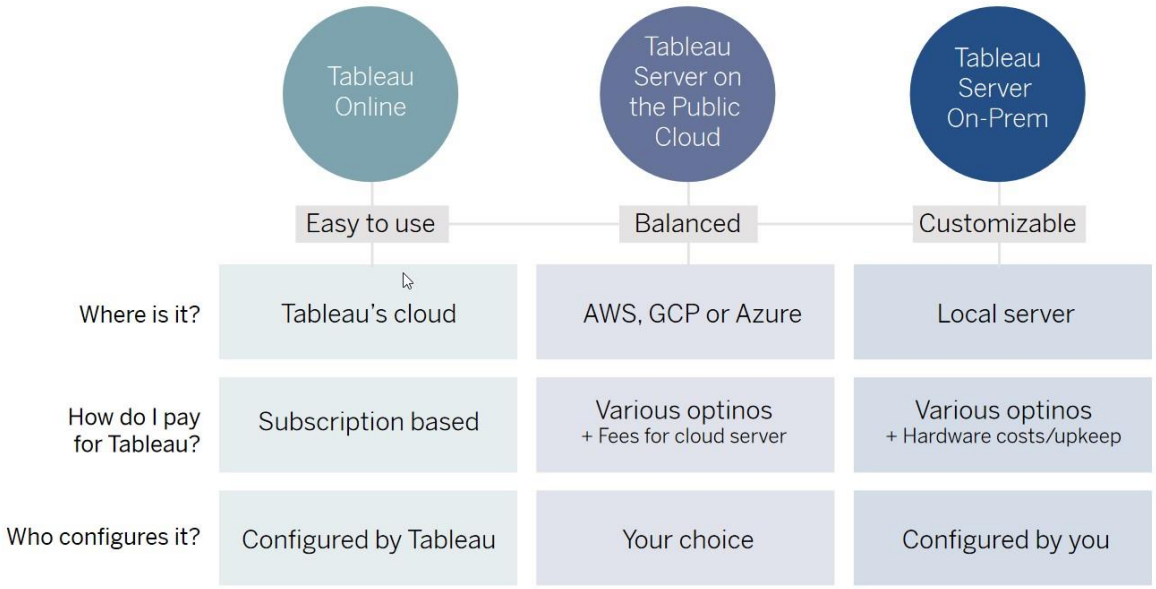
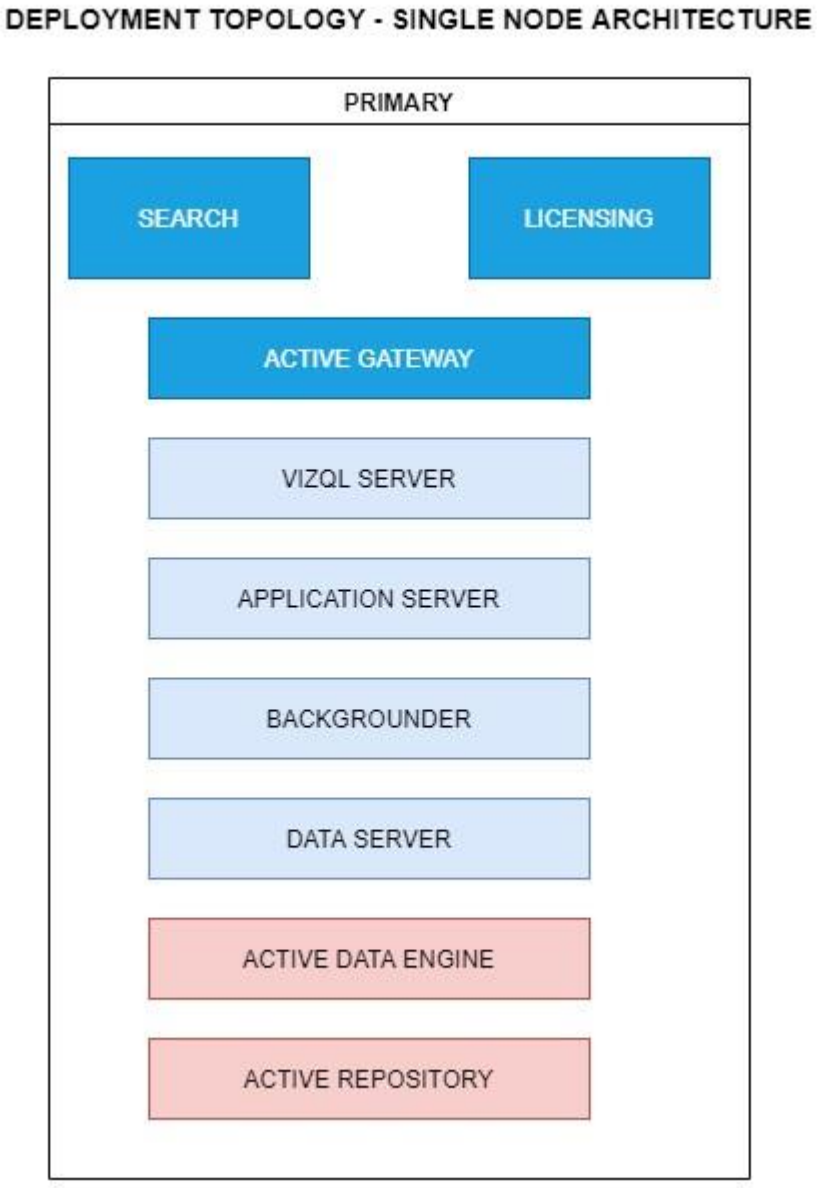


Tableau Online: Get up and running quickly with no hardware required. Tableau Online is fully hosted by Tableau so all upgrades and maintenance are automatically managed for you.

Tableau Server deployed on public cloud: Leverage the flexibility and scalability of cloud infrastructure without giving up control. Deploy to Amazon Web Services, Google Cloud Platform, or Microsoft Azure infrastructure to quickly get started with Tableau Server (on your choice of Windows or Linux). Bring your own license or purchase on your preferred marketplace.

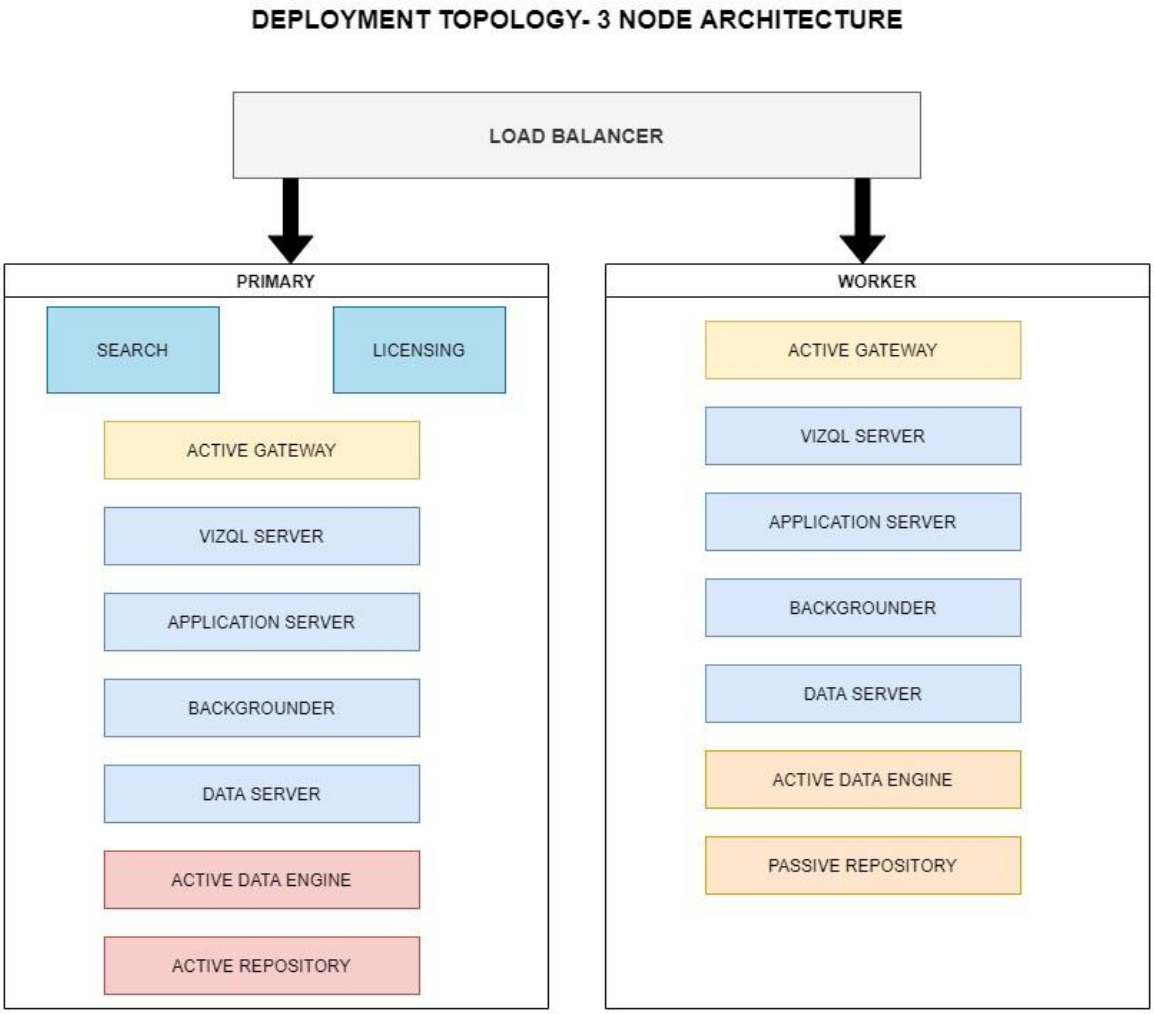
Tableau Server deployed on-premises: Manage and scale your own hardware and software (whether Windows or Linux) as needed. Customize your deployment as you see fit.

3.2 Single Node Architecture



This architecture is a single node architecture. This is the most simple deployment topology.

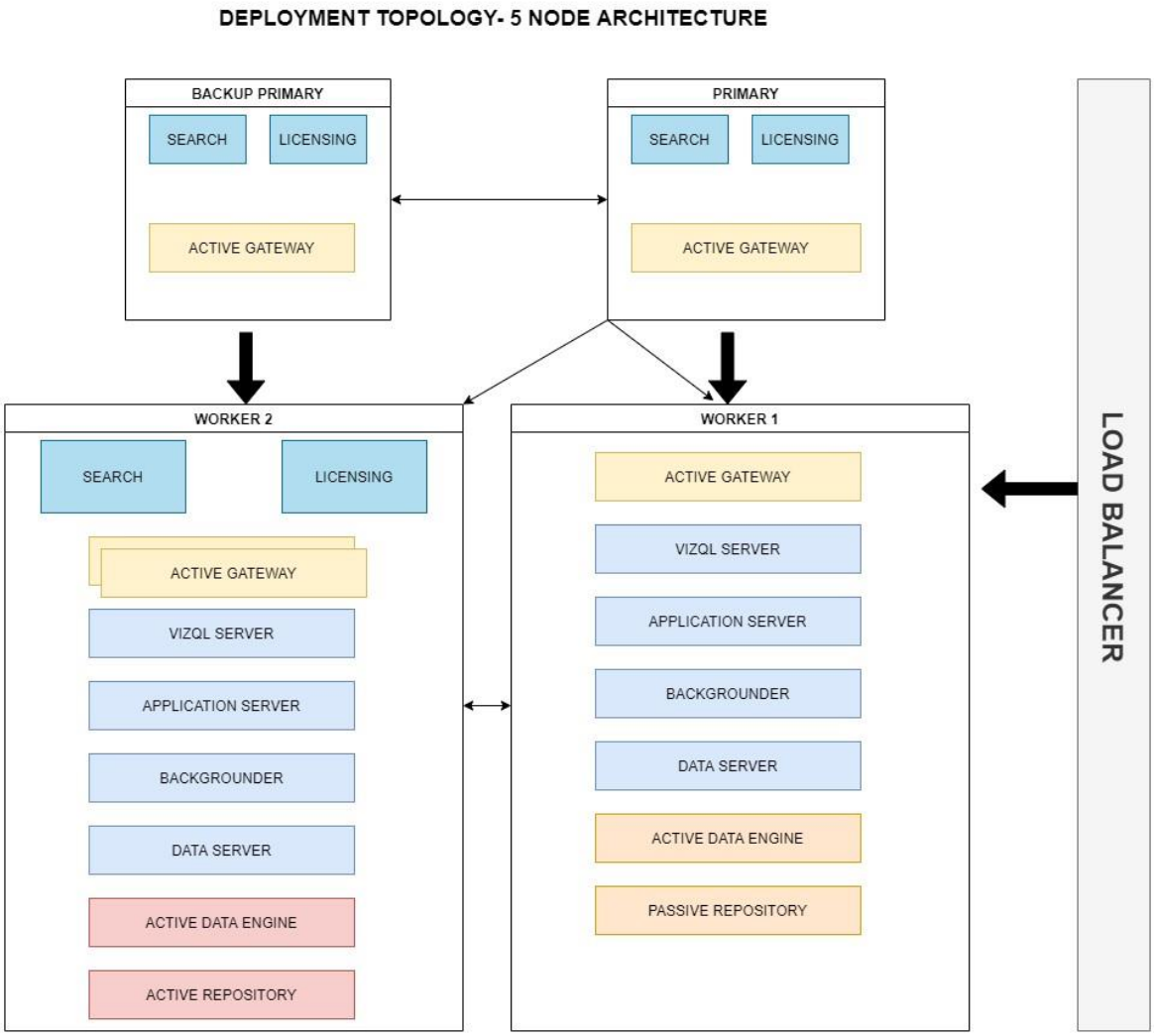
3.3 3 Node Architecture



This architecture is a 3 Node Architecture which is more capable to handle concurrent requests.

If we need failover or high availability, or want a second instance of the repository, we must install Tableau Server on a cluster of at least three computers. In a cluster that includes at least three nodes, you can configure two instances of the repository, which gives our cluster failover capability.

3.4 5 Node Architecture



When we install Tableau Server on a Five-node cluster, we can install server processes on one or both nodes. A five-node cluster can improve the performance of Tableau Server, because the work is spread across multiple machines.

Note the following about five-node clusters:

- A five-node cluster does not provide failover or support for high availability.
- You can't install more than one instance of the repository on a two-node cluster, and the repository must be on the initial node.