Architecture Design

Amazon Sales Data Analysis

Written By	Dipankar Modak, Akash Sahu
Document Version	0.1
Last Revised Date	

DOCUMENT CONTROL

Change Record:

VERSION	DATE	AUTHOR	COMMENTS

Reviews:

VERSION	DATE	REVIEWER	COMMENTS

Approval Status:

VERSION	REVIEW DATE	REVIEWED BY		APPROVED BY	COMMENTS

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.

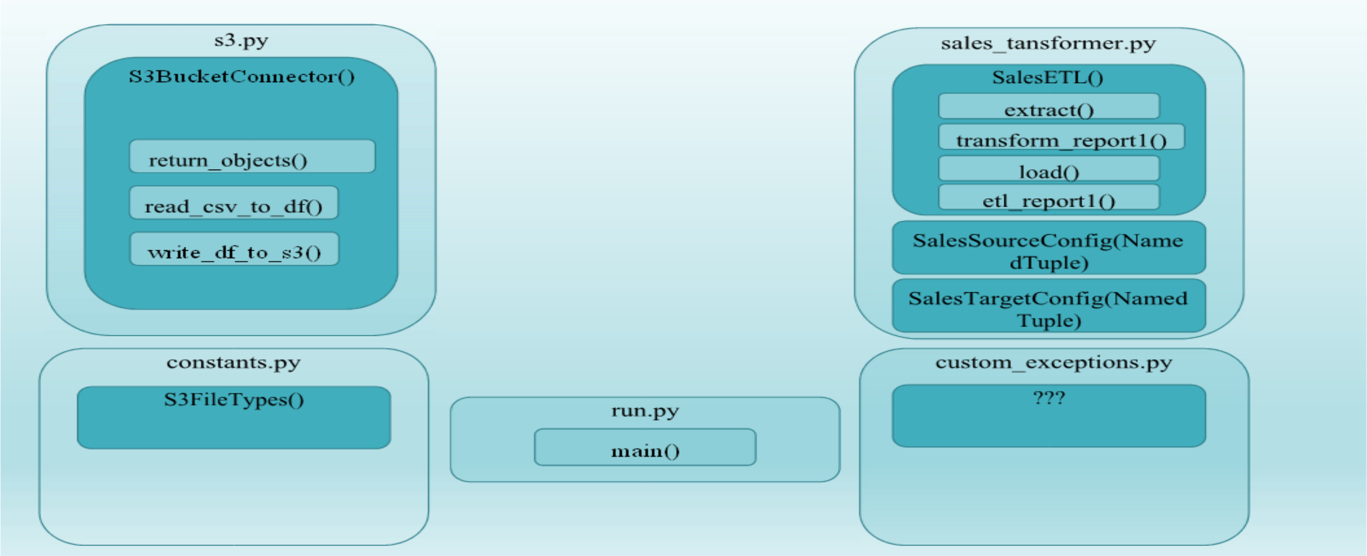
Each style will describe a system category that consists of :

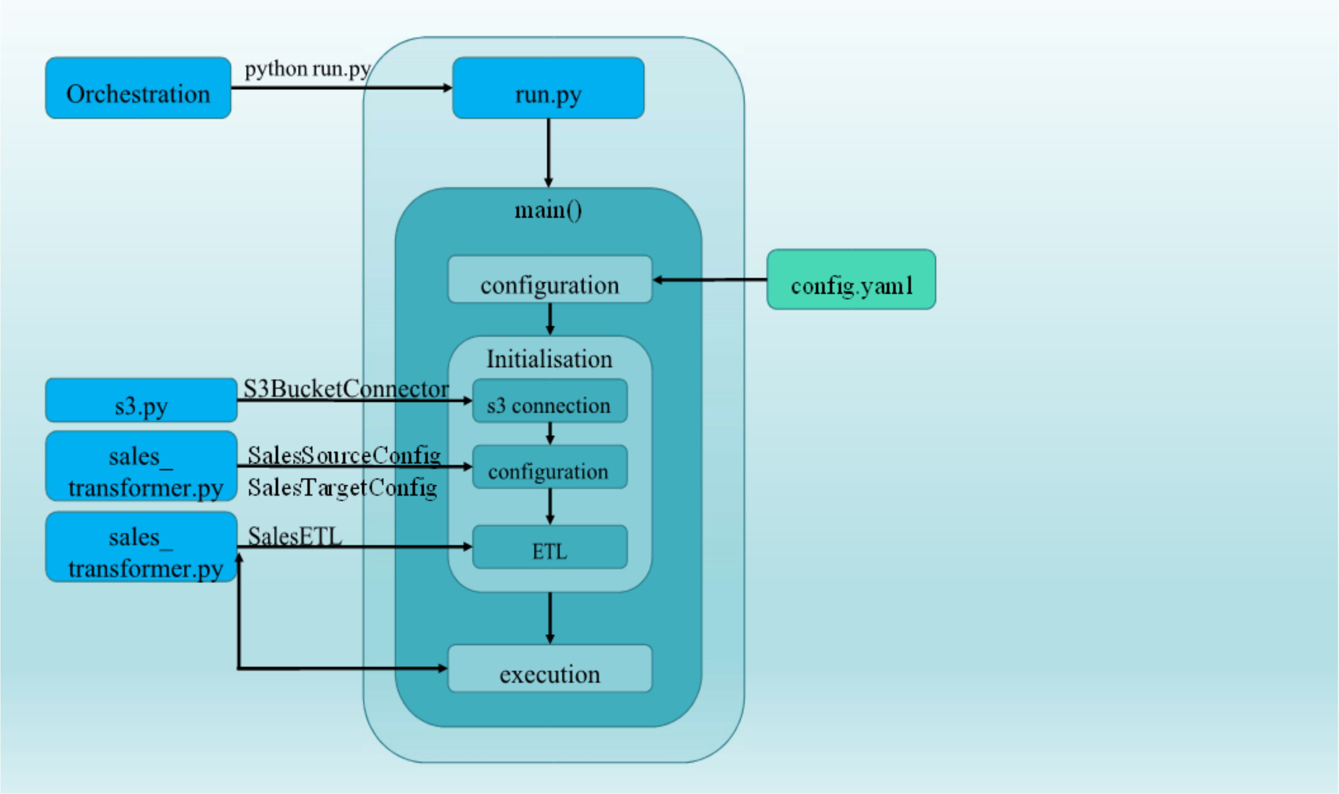
- A set of components (eg: a database, computational modules) that will perform a function required by the system.
- The set of connectors will help in coordination, communication, and cooperation between the components.
- Conditions that how components can be integrated to form the system.
- Semantic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

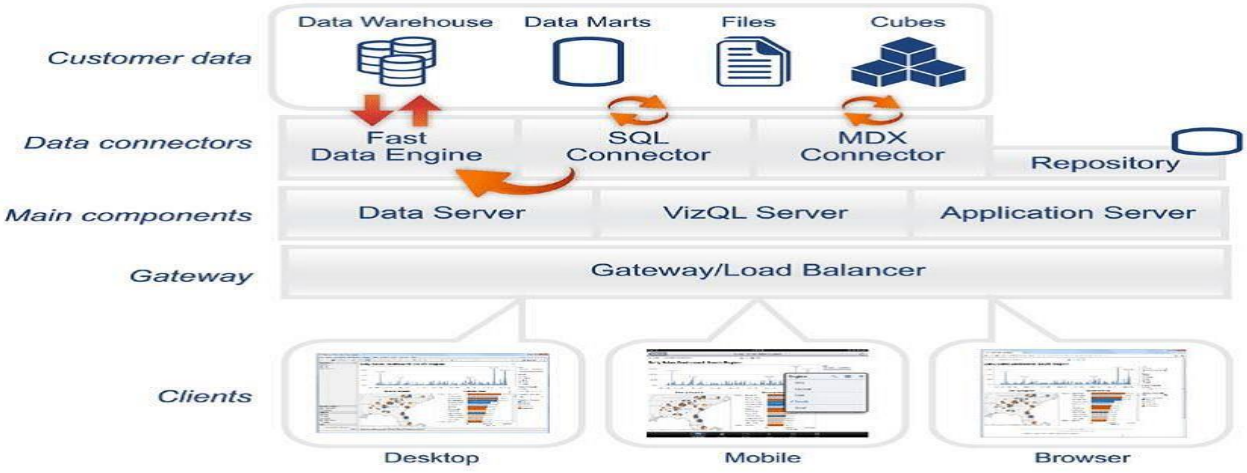
2. Architecture

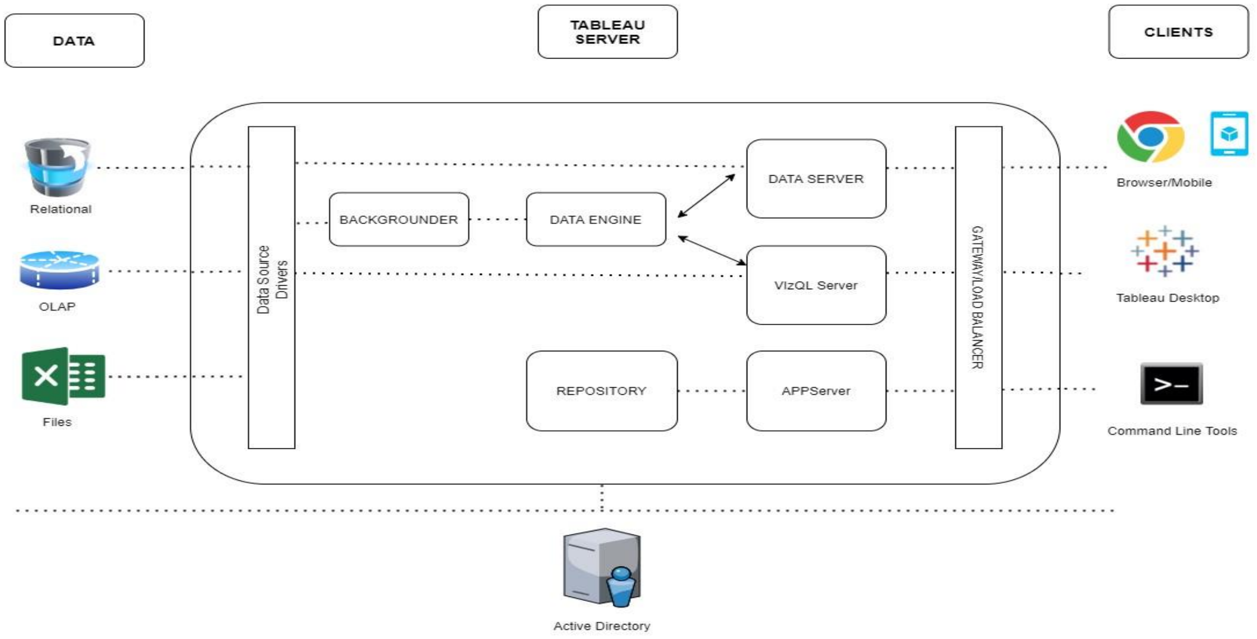




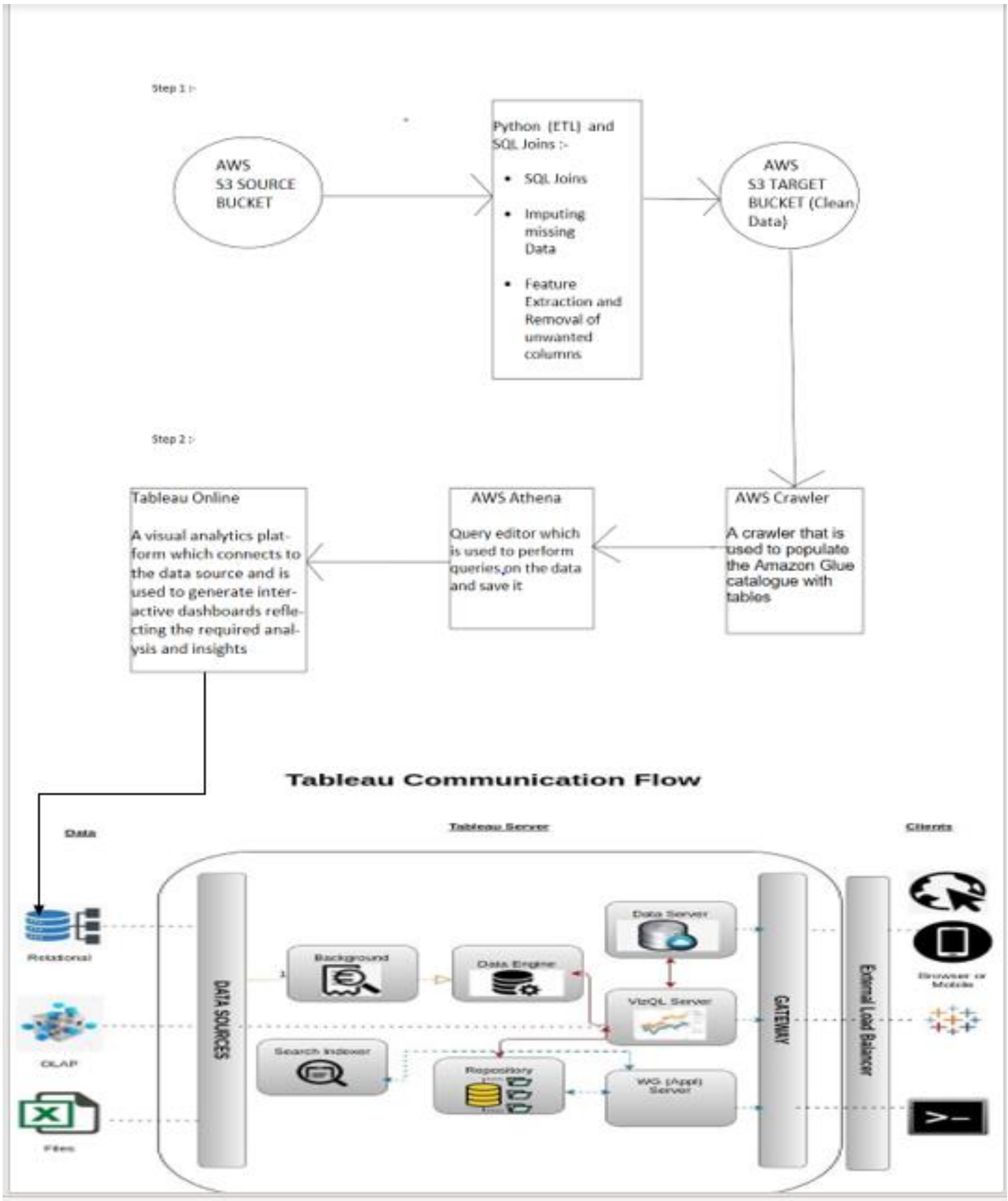
The ETL pipeline created in python

The following diagrams shows Tableau Server’s architecture:





Total System Architecture



- SQL Joins
- Imputing missing Data
- Feature Extraction and Removal of unwanted columns

A visual analytics platform which connects to the data source and is used to generate interactive dashboards reflecting the required analysis and insights

Query editor which is used to perform queries on the data and save it

A crawler that is used to populate the Amazon Glue catalogue with tables

ETL -Pipeline

2.2 AWS S3 bucket

The ETL pipeline consists of remote data storage services like AWS S3 bucket. Amazon S3 buckets, which are similar to file folders, store objects, which consist of data and its descriptive metadata. Python is used to write the ETL methods which then interacts with the S3 buckets using Boto3 library.

Boto3 is the name of the Python SDK for AWS. It allows you to directly create, update, and delete AWS resources from your Python scripts.

2.3 AWS Crawler

A crawler is a job defined in Amazon Glue. It crawls databases and buckets in S3 and then creates tables in Amazon Glue together with their schema. Then, you can perform your data operations in Glue, like ETL. Athena can connect to your data stored in Amazon S3 using the AWS Glue Data Catalog to store metadata such as table and column names. After the connection is made, your databases, tables, and views appear in Athena's query editor.

2.4 AWS Athena

1. Amazon Athena is an interactive query service that makes it easy to analyze data in Amazon S3 using standard SQL. Athena is serverless, so there is no infrastructure to manage, and you pay only for the queries that you run.

2. Athena is easy to use. Simply point to your data in Amazon S3, define the schema, and start querying using standard SQL. Most results are delivered within seconds. With Athena, there's no need for complex ETL jobs to prepare your data for analysis. This makes it easy for anyone with SQL skills to quickly analyse large-scale datasets.

3. Athena is out-of-the-box integrated with AWS Glue Data CataLog, allowing you to create a unified metadata repository across various services, crawl data sources to discover schemas and populate your Catalog with new and modified table and partition definitions, and maintain schema versioning.

2.5 ETL Pipeline Script

Extract Transform and Load Workflow that is written in Python. The config. yaml file contains all the configuration settings that are needed to deploy the Pipeline. From the config. yaml file, you can customize your installation by using various parameters. This makes the source code more effective and more readable.The Clean dataset is load to S3 bucket

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.

Tableau Server is internally managed by the multiple server processes.

2.6 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.

2.7 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.

2.8 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.

2.9 VIZQL Server:-

Once a view is opened, the client sends a request to the VizQL process (vizqlserver.exe). The VizQL process then sends queries directly to the data source, returning a result set that is rendered as images and presented to the user. Each VizQL Server has its own cache that can be shared across multiple users

2.10 Data Engine:-

It Stores data extracts and answers queries.

2.11 Backgrounder:-

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

2.12 Data Server:-

Data Server Manages connections to Tableau Server data sources

It also maintains metadata from Tableau Desktop, such as calculations, definitions, and groups.

2.13 Tableau Communication Flow

3. Deployment Description

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:

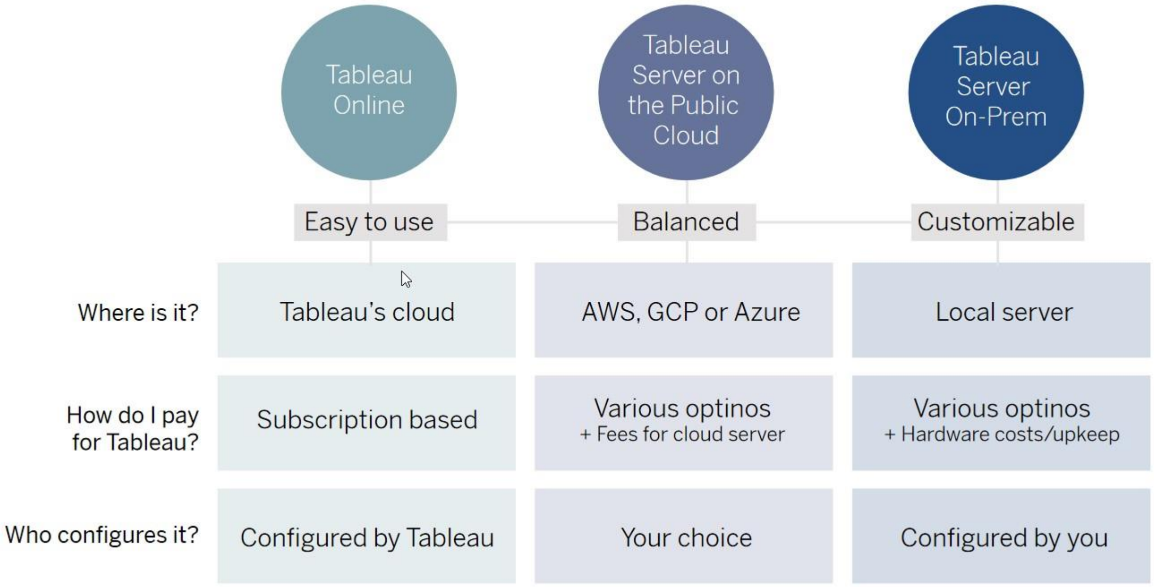
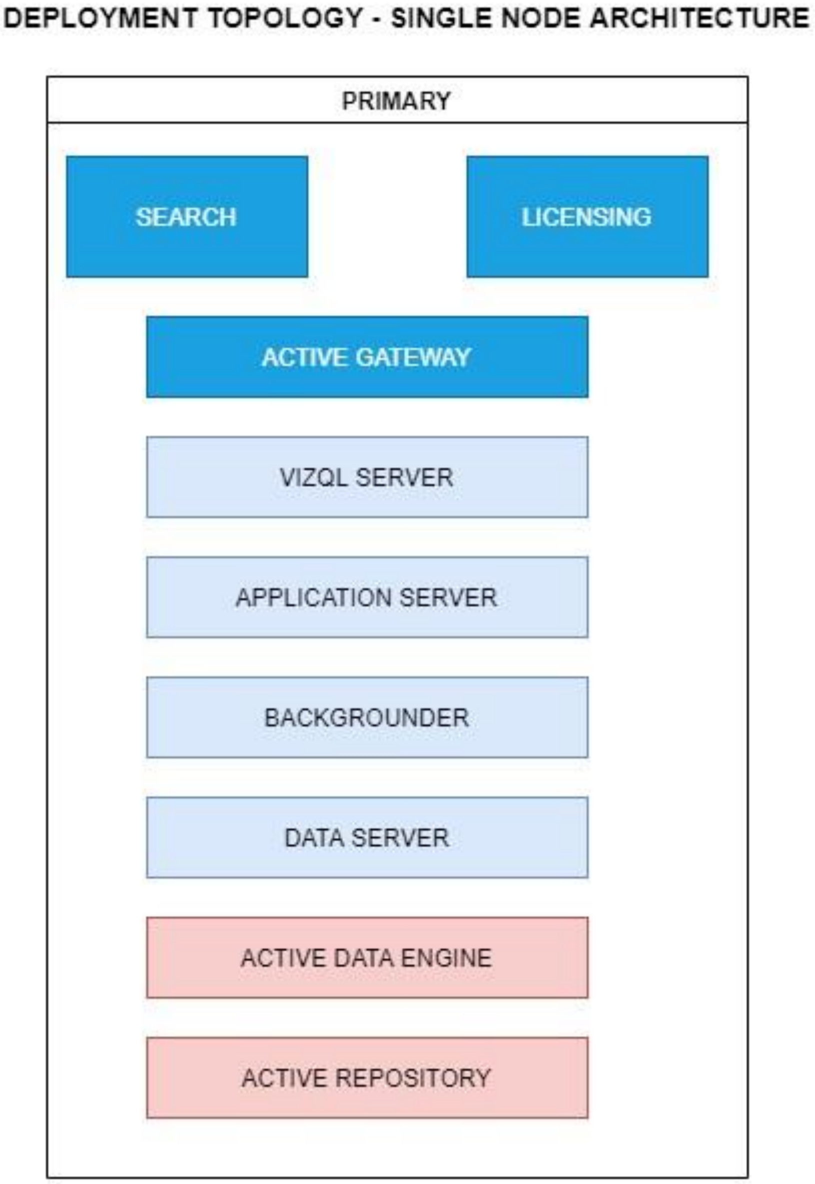


	Tableau Online	Tableau Server on the Public Cloud	Tableau Server On-Prem
	Easy to use	Balanced	Customizable
Where is it?	Tableau's cloud	AWS, GCP or Azure	Local server
How do I pay for Tableau?	Subscription based	Various optinos + Fees for cloud server	Various optinos + Hardware costs/upkeep
Who configures it?	Configured by Tableau	Your choice	Configured by you

1. **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.
2. **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.
3. **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.

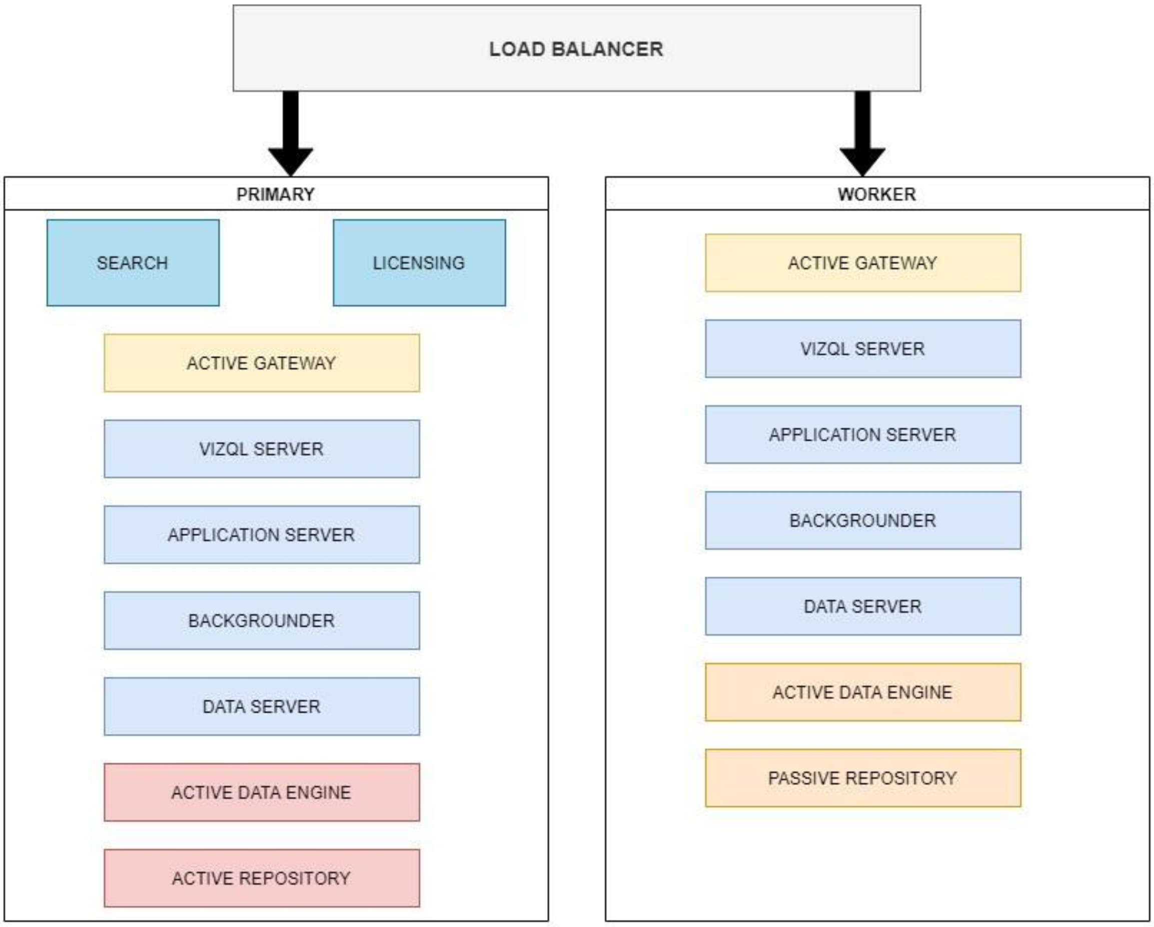
3.2 Single Node Architecture



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

3.3 3 Node Architecture

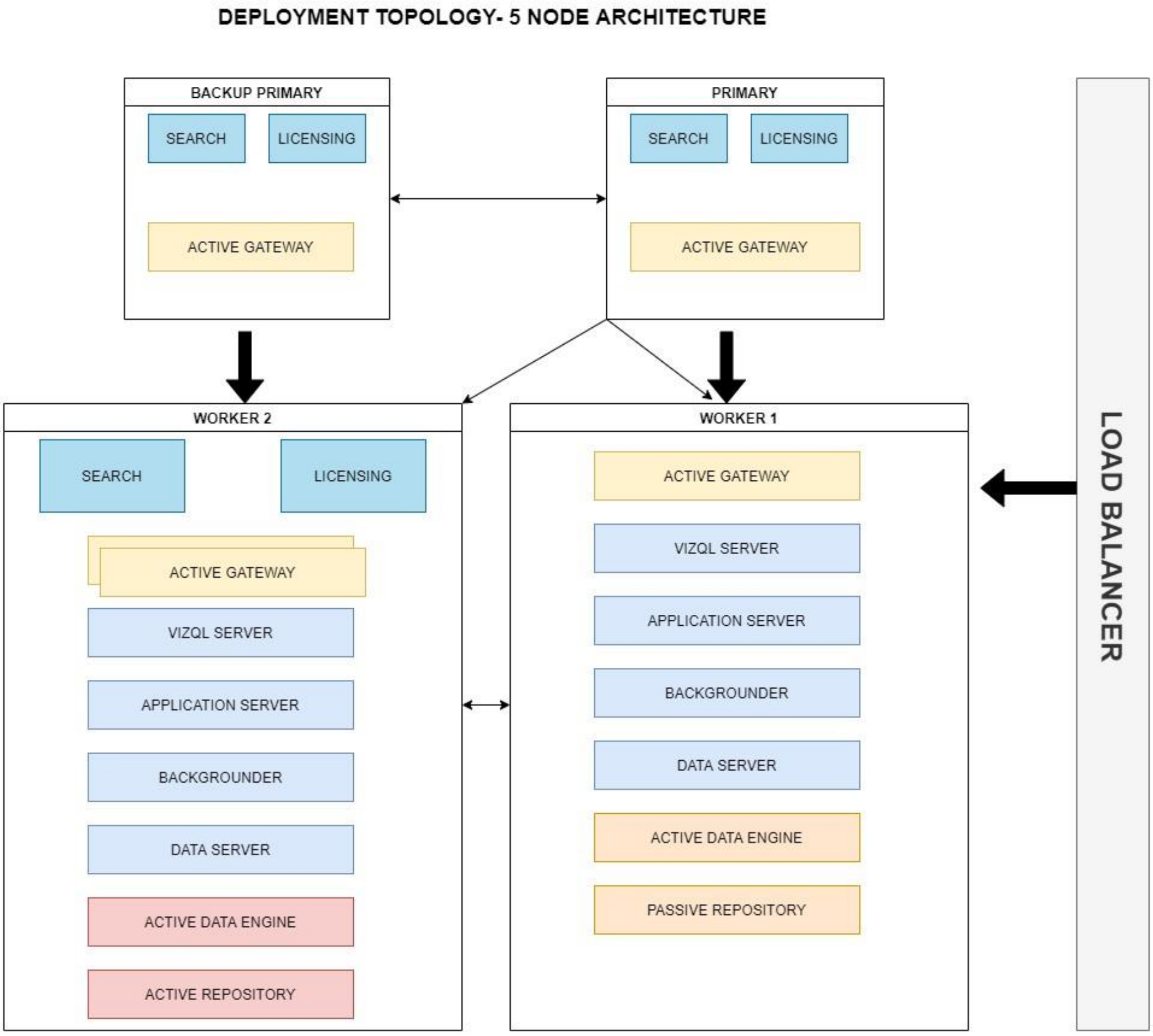
DEPLOYMENT TOPOLOGY- 3 NODE ARCHITECTURE



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.