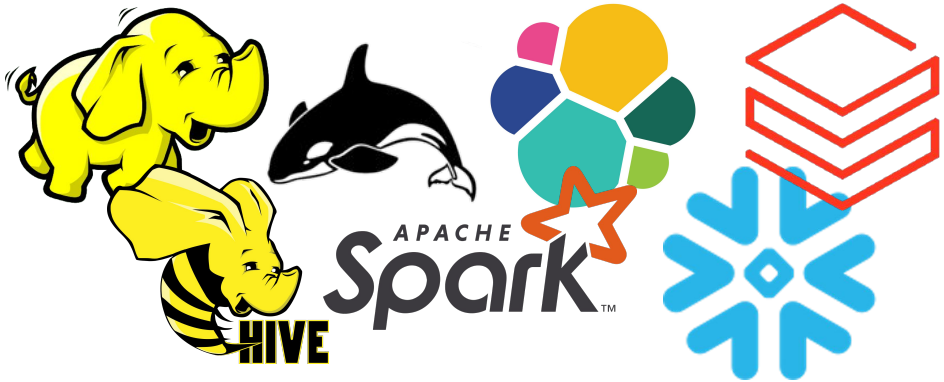
ADALTAS

Big Data Ecosystem



Data Governance

Challenges

The more data, the more **challenges** organizations meet:

- What data assets do we have?
- Who is the owner?
- How is it transformed?
- What is the data quality?
- Who has an access?
- What data is private (GDPR)?
- ...

Possible use cases

1. Customer names may be listed differently in sales, logistics and customer service systems

   => evolves data integrity issues

2. An available data asset is unknown for data scientists

   => missed opportunity to extract value

Data Governance

It is a **set of practices** to ensure important data are formally managed through the company to achieve:

- availability
- usability
- integrity
- security

Data Governance initiatives

- Data catalog
- Data mapping and classification
- Data lineage
- Audit logs
- Business glossary
- Privacy and security
- ...

Data catalog

- collect metadata from systems and use it to create an indexed inventory of available data assets
- includes information on data lineage, search functions and collaboration tools

Data mapping and classification

- helps document data assets
- defines how data flows through an organization
- classification based on factors (contain personal information or other sensitive data)
- influence how data governance policies are applied

Business glossary

- contains definitions of business terms and concepts used in an organization
- Eg: what constitutes an “active customer”

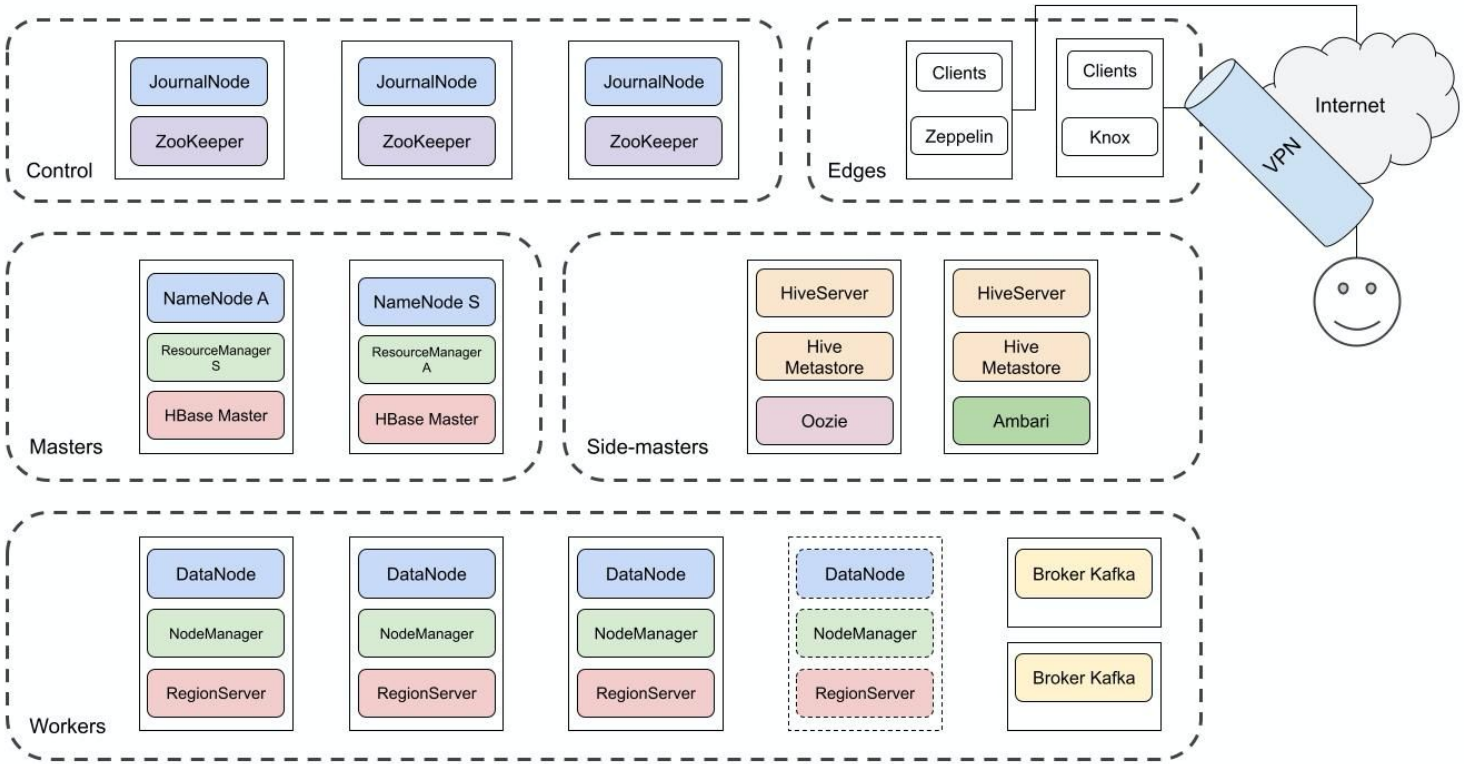
Data lineage

- **the journey** data takes from its creation through its transformations over time
- Answers to:
  - Who created data?
  - How it is transformed?

Apache Atlas

- Data Governance framework
- https://atlas.apache.org
- Provides:
  - metadata management
  - classify and govern
  - collaboration capabilities

Hadoop cluster topology



Hadoop cluster topology

Node types:

- **Masters:** NN, RM, HBaseMaster
- **Utility nodes:** HiveMetastore, Oozie, Ambari
- **Workers:** DN, NM, RS
- **Edge nodes:** HS2, clients (hdfs, yarn, beeline, hbase, spark)
- **Security nodes**

Hadoop cluster topology

Node hardware specifications:

- **Masters:** medium RAM/CPU, **RAID** on disks
- **Side-masters:** medium RAM/CPU
- **Workers:** lot of RAM/CPU, lots of disks (> 10), no RAID
- **Edge nodes:** can be VMs/containers
- **Security nodes**

Security

3 main principles:

- **Identification:** indicate user's identity
- **Authentication:** prove the user's identity (e.g. password)
- **Authorization:** check user's access rights to resources
- **+ Privacy** = Encryption

Security: locally

Unix permissions (in Linux, MacOS):

- **UID + GID** (User ID, Group ID)
- Identification only
- **Security holes:** possible to impersonate a user by matching the UID/GUID → e.g. HDFS client running in a container

Identification: LDAP

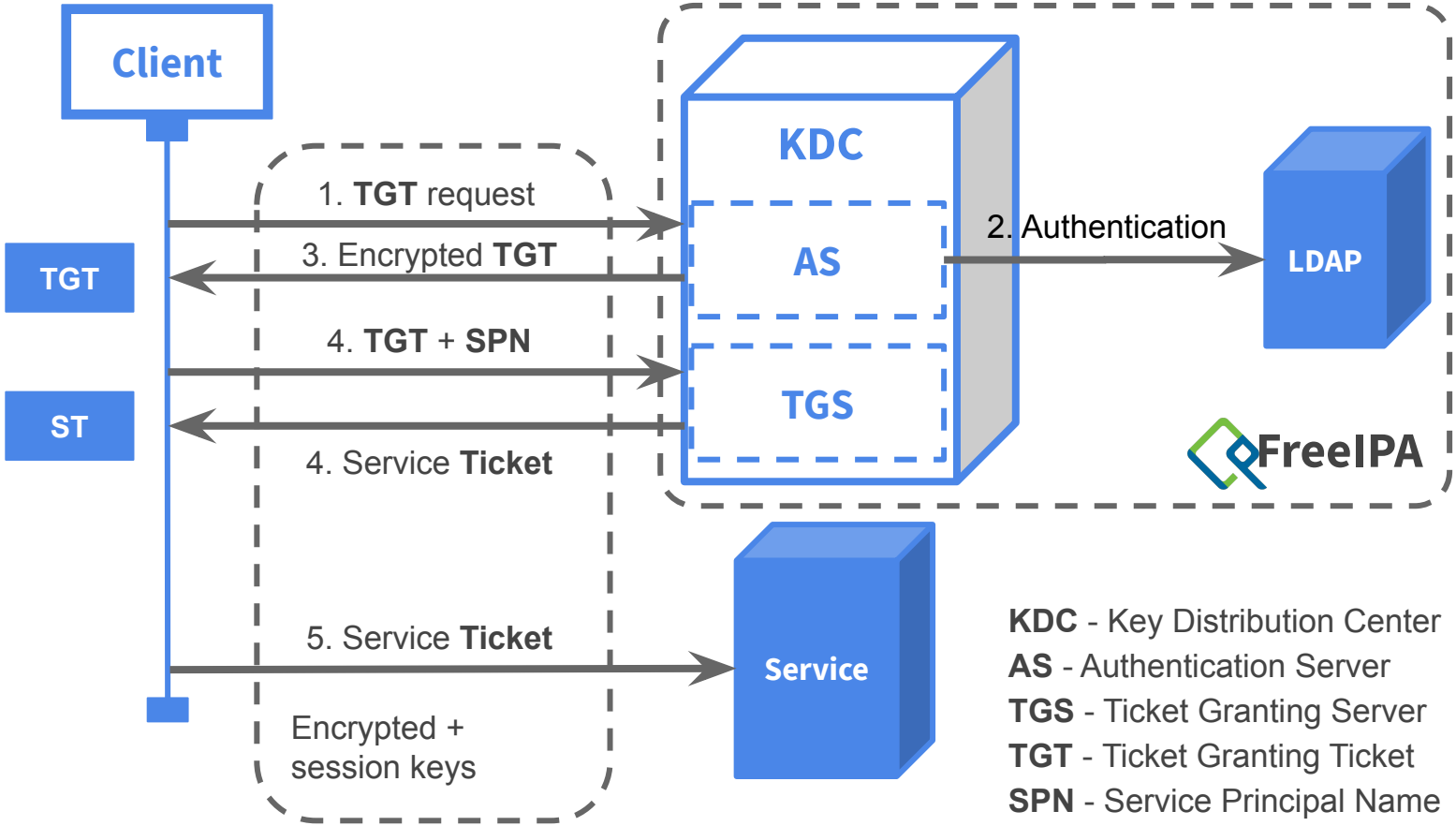
LDAP = Lightweight Directory Access Protocol

- Stores **users** and **groups**
- Allows **identification** (“this user exists and belongs to those groups”)
- Also stores passwords for basic authentication

Examples: OpenLDAP, FreeIPA, Active Directory

Authentication: Kerberos

- Authentication based on a **ticketing system**
- Single Sign-On (SSO)
- **Control access to services** by authenticating the users



https://en.wikipedia.org/wiki/Kerberos_(protocol)#Protocol

Authorization: Apache Ranger

RBAC (Role Based Access Control) on Hadoop:

- HDFS (*rwx* on folders)
- YARN (access to queues)
- Hive (access to tables, columns)
- HBase (access to tables, column families, columns)

Integration with LDAP

Privacy: Encryption in Hadoop

- Possible usage of **SSL/TLS** (like HTTPS) for services and client-service communications
- Wire encryption
- Encryption at rest
- Performance impact