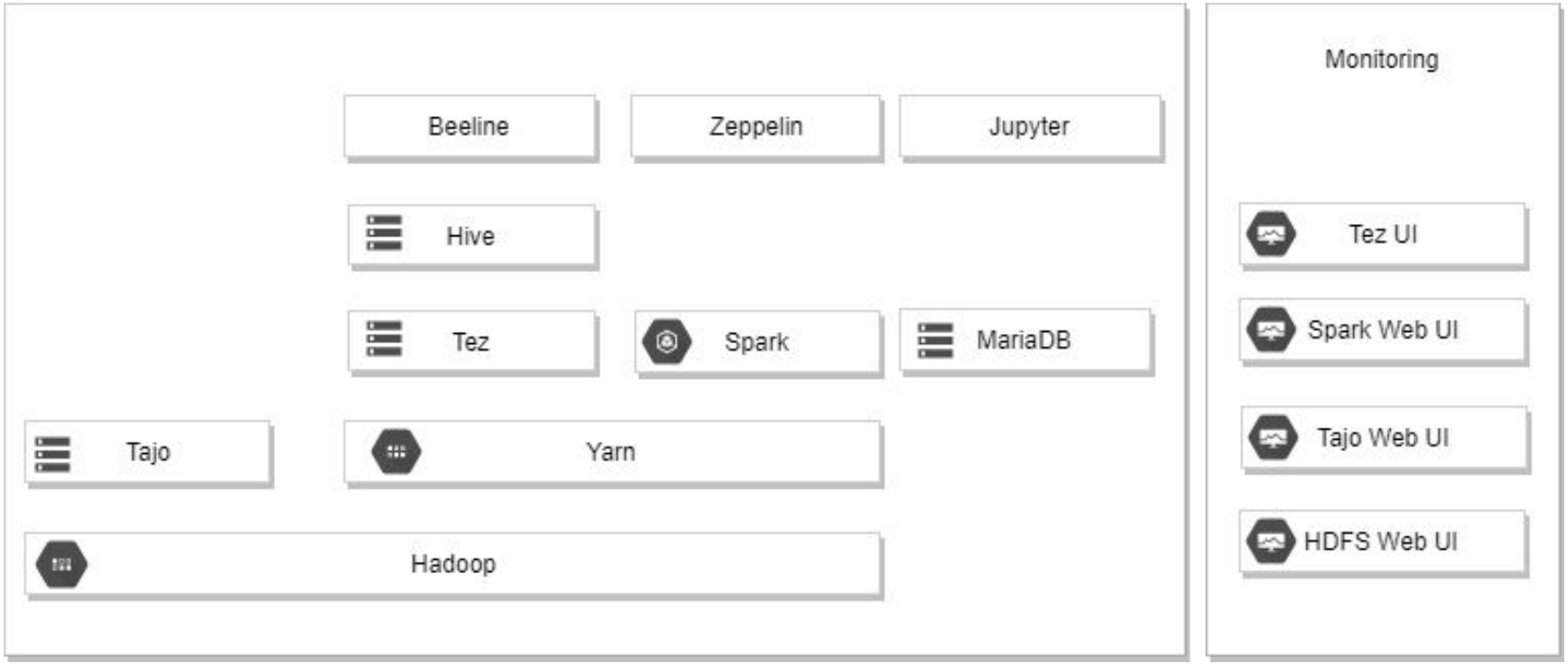
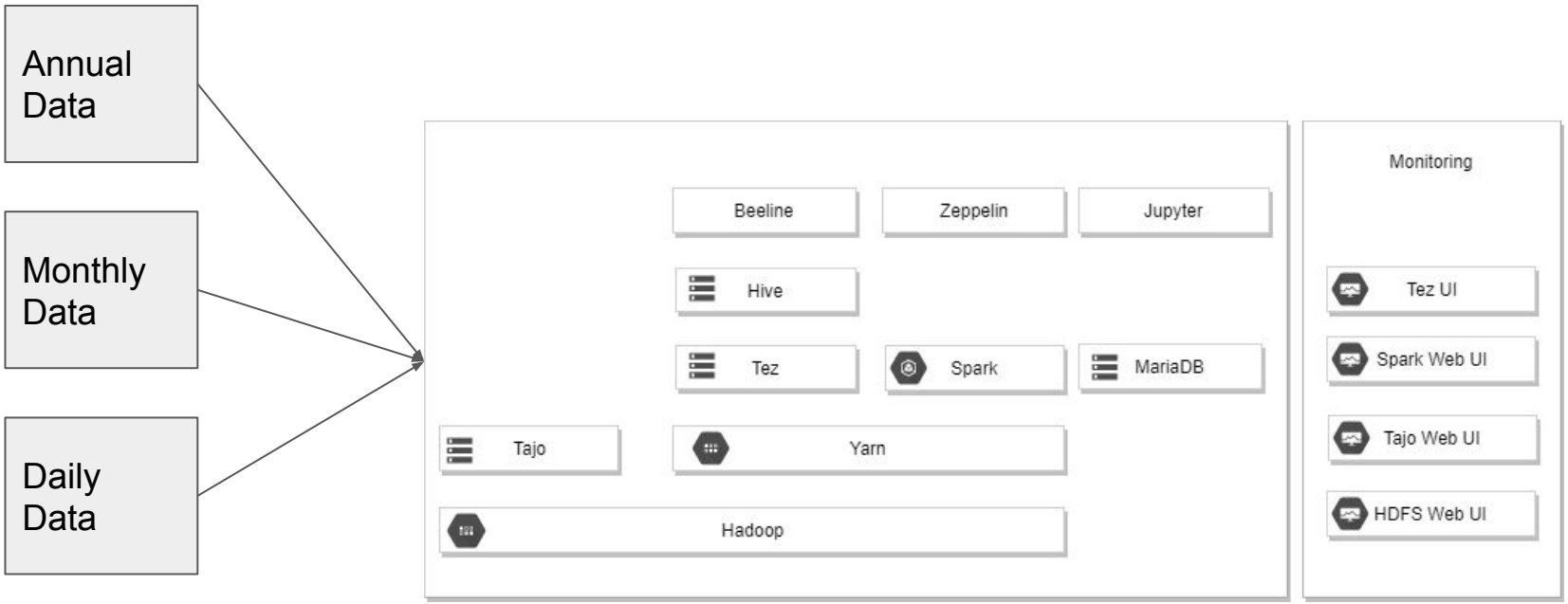
Lightning Talk:
Hive & Spark & Zeppelin

Cinyoung Hur
Seoul AI
2017-12-09

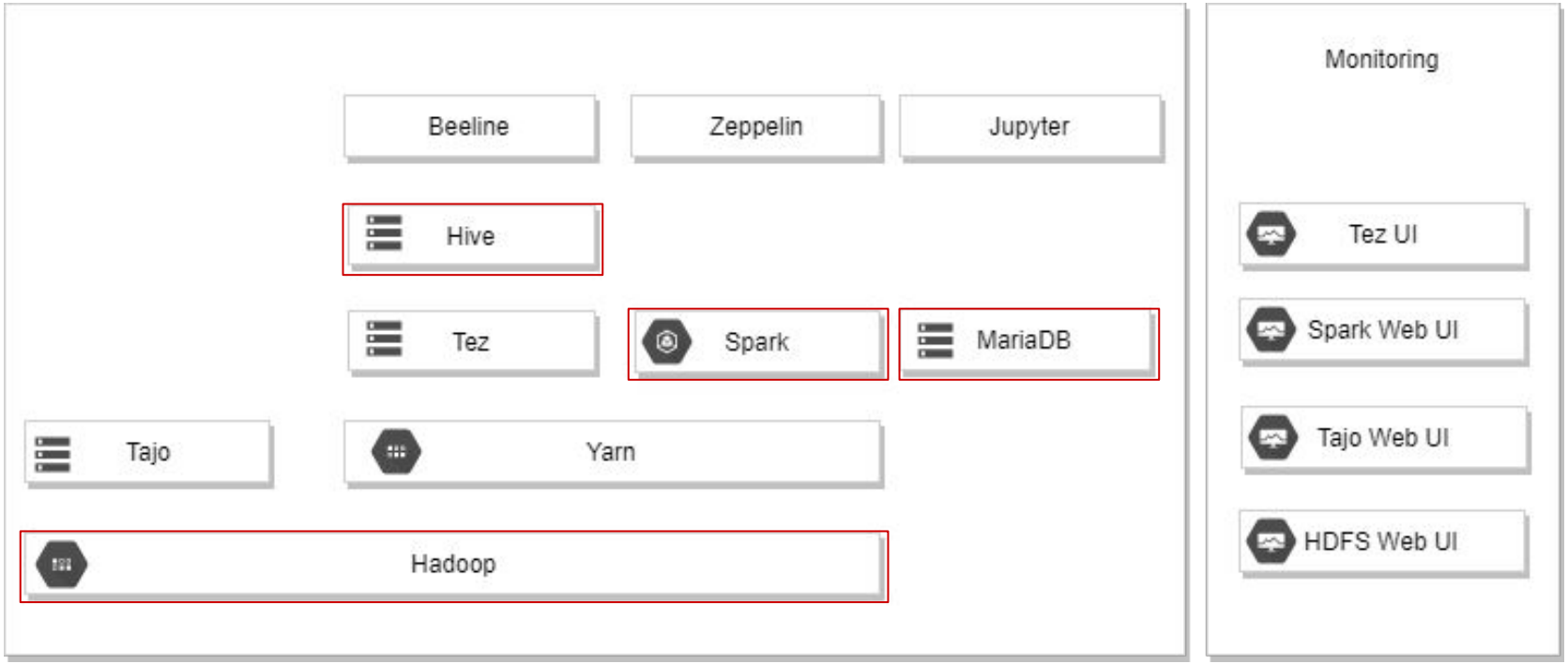
Data Lake architecture



Collecting Data



Dynamically joining data frames

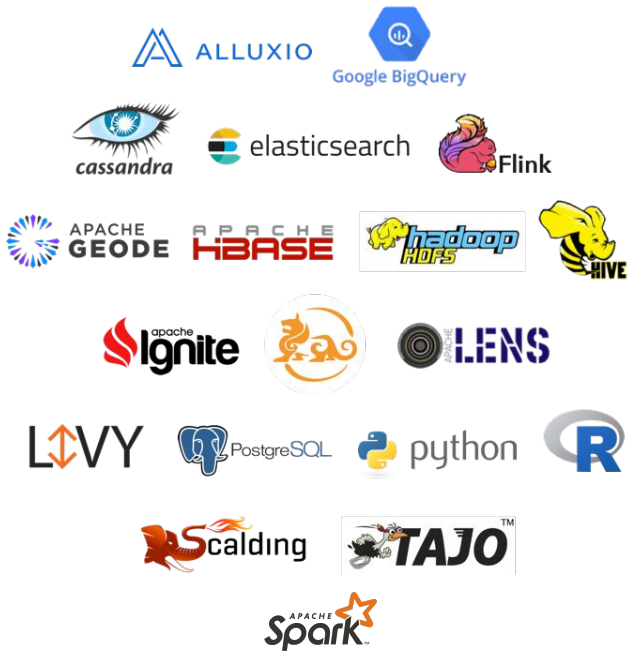


DataFrames

- A DataFrame is a DataSet organized into named columns
  - Like table in a relational databases
- DataFrames can be constructed from various sources
  - Structured data files(json, csv), tables in Hive, external databases, existing RDDs

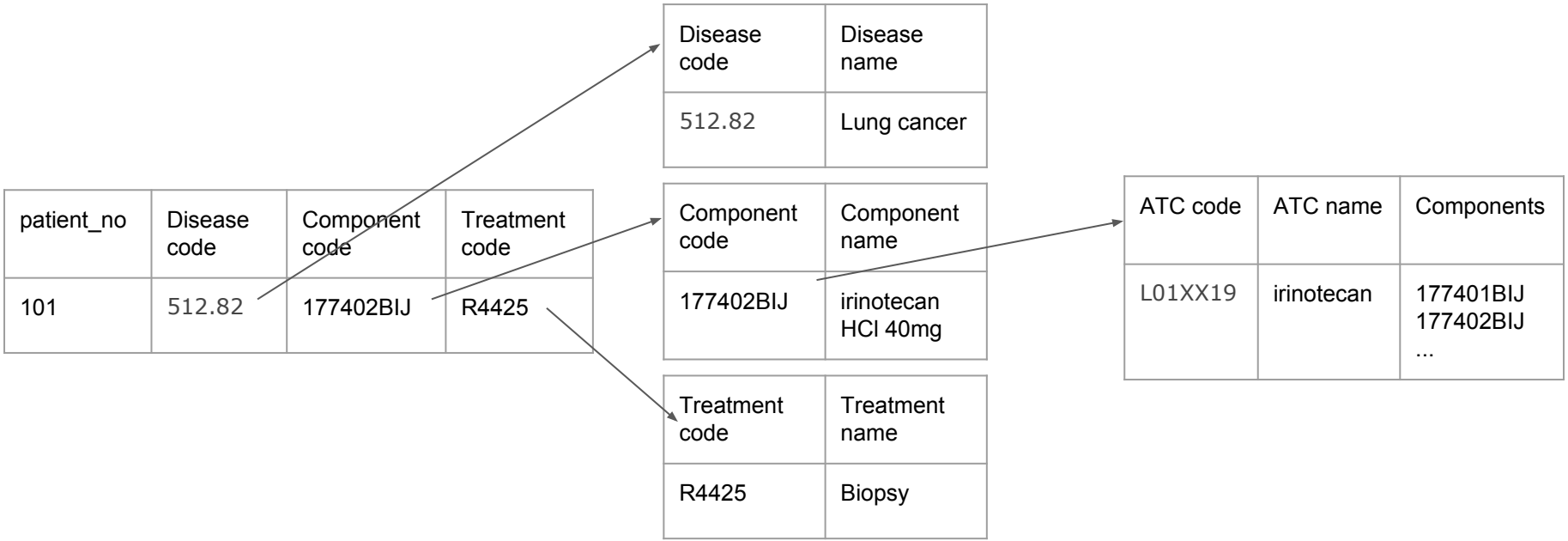
Apache Zeppelin

- Multi-purpose Notebook
- Useful features
  - Multi-user support(LDAP)
  - Built-in version control
  - Collaboration and Personal mode
  - Pluggable visualization
  - Various interpreters
  - … and much more



Joining DataFrames in Zeppelin

Use case: medical claims & its metadata in MariaDB & third-party source in a csv format



Joining DataFrames in Zeppelin

```
%pyspark
atcInfo = spark.read \
        .format("jdbc") \
        .option("url", "jdbc:mysql://DB_URL:3306/DBNAME") \
        .option("driver", "com.mysql.jdbc.Driver") \
        .option("dbtable", "ComponentATCInfo") \
        .option("user", "user") \
        .option("password", "password") \
        .load()
atcInfo.registerTempTable("atcInfo")

diseaseInfo = spark.read \
    .format("jdbc") \
    .option("url", "jdbc:mysql://DB_URL:3306/DBNAME") \
    .option("driver", "com.mysql.jdbc.Driver") \
    .option("dbtable", "DiseaseInfo") \
    .option("user", "user") \
    .option("password", "password") \
    .load()
diseaseInfo.registerTempTable("diseaseInfo")
```

Joining DataFrames in Zeppelin

```
%pyspark
from pyspark.sql import HiveContext
from pyspark.sql.functions import *

hiveCtx = HiveContext(sc)

mc = hiveCtx.table("medical_claims_2016")
mc = mc.join(atcInfo, mc.compoment_code == atcInfo.code, how='inner')
mc = mc.join(diseaseInfo, mc.disease_code == diseaseInfo.code, how='inner')
mc.registerTempTable("mc_extended")
```

patient_no	Disease code	Disease name	Component code	Treatment code	Component name
101	512.82	Lung cancer	177402BIJ	R4425	irinotecan HCl 40mg