Data Science with Apache Zeppelin

DEVIEW
2015

이문수
NFLabs

contents

DEVIEW
2015

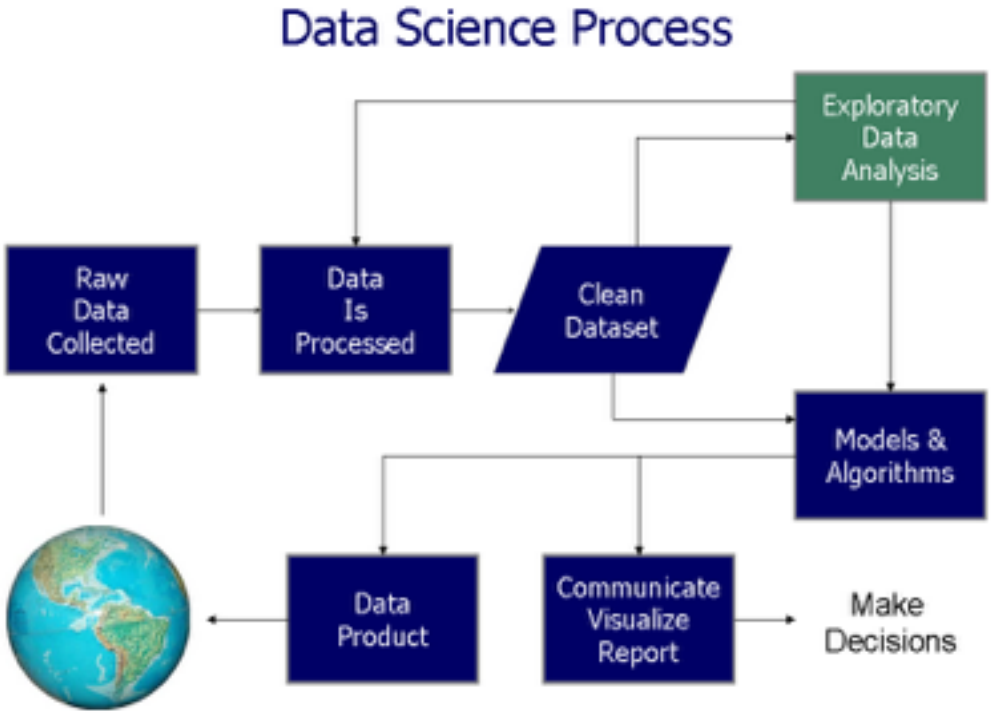
1. Data science lifecycle
2. Apache Zeppelin
3. Zeppelin in your team
4. Helium

DEVIEW
2015

1.
Data Science Lifecycle

Data Science

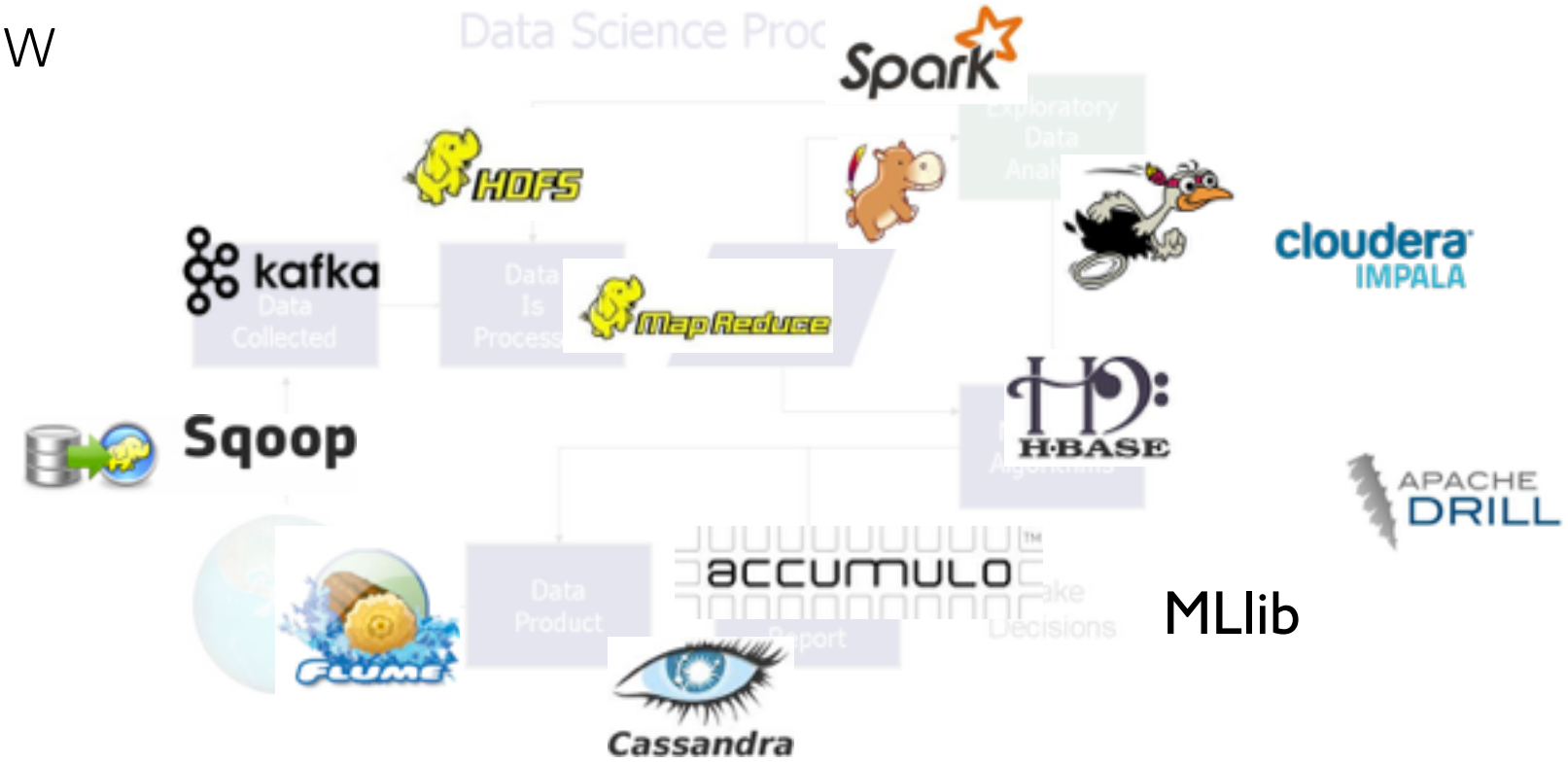
다양한 과정



https://en.wikipedia.org/wiki/Data_analysis

Data Science

다양한 SW



Data Science

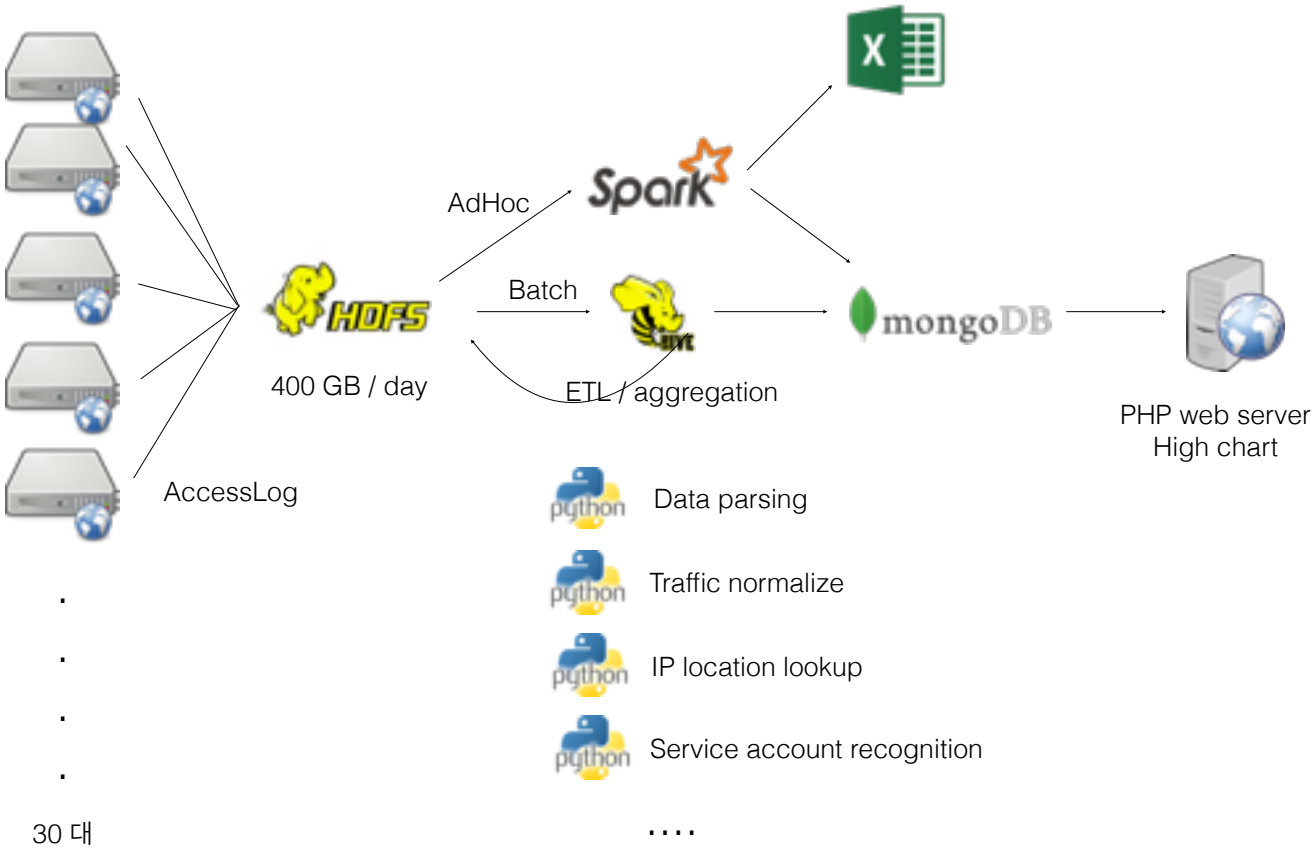
DEVIEW
2015

다양한 사람

즉

힘든 삶

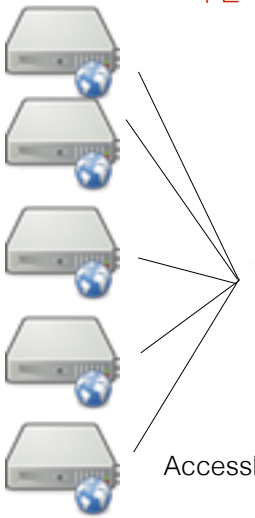
힘든 삶


AdHoc
Spark
Batch
HDFS
400 GB / day
ETL / aggregation
mongoDB
PHP web server
High chart
AccessLog
Data parsing
Traffic normalize
IP location lookup
Service account recognition
30 대
....

Spark 에서 AdHoc query를 위해 Hive 에서 aggregation 을 미리 수행해야 하는 경우가 많은데, 소스가 따로 떨어져있어서 관리가 잘안되고 실수가 자주 발생

Excel 로 리포트 만들어 이메일 보내면 매일 매일 업데이트 해달라고 연락옴

데이터 넣는 과정에서 오류 발생하면 수동으로 재처리



AccessLog

30 대

HDFS

400 GB / day

AdHoc

Spark

Batch

Hive

ETL / aggregation

mongoDB

PHP web server
High chart

- Data parsing
- Traffic normalize
- IP location lookup
- Service account recognition
- ....

새로운 분석 추가될때 마다 mongodb 에 스키마 관리 웹페이지 수정 및 차트 추가

분석에 집중할 수 없을까?

DEVIEW
2015

적절한 도구가 필요

편리하고 강력한 분석 언어
인터렉티브 속도
라이브러리
시각화
공유/협업
간편하고 손쉬운 시스템 구성

2.
Apache Zeppelin

Apache Zeppelin

http://zeppelin.incubator.apache.org/

2014 년 12월에 ASF 에 incubation 됨

2013 년 8월에 NFLabs 내부 프로젝트로 시작

63 Contributors from worldwide

646 Stars

1 release

Apache 2.0 License

누가 쓰나?


at&t
NAVER
SK
B
NETFLIX
IBM
Microsoft
intel
GE
ebay
Spotify
LINE
cash slide
@WalmartLabs
TOYOTA
Y!
SAMSUNG
trulia
BOSCH
Bluehole studio

생태계

Collaboration, Sharing

ZeppelinHub

Packaging & Deployment

Hortonworks

Ambari

Z-Manager

Zeppelin + Full stack on a cloud

Microsoft Azure

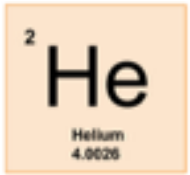
SEQUENCEIQ

Qubole

juju

DATALAYER

Packages



Backend integration

Spark

Flink

HIVE

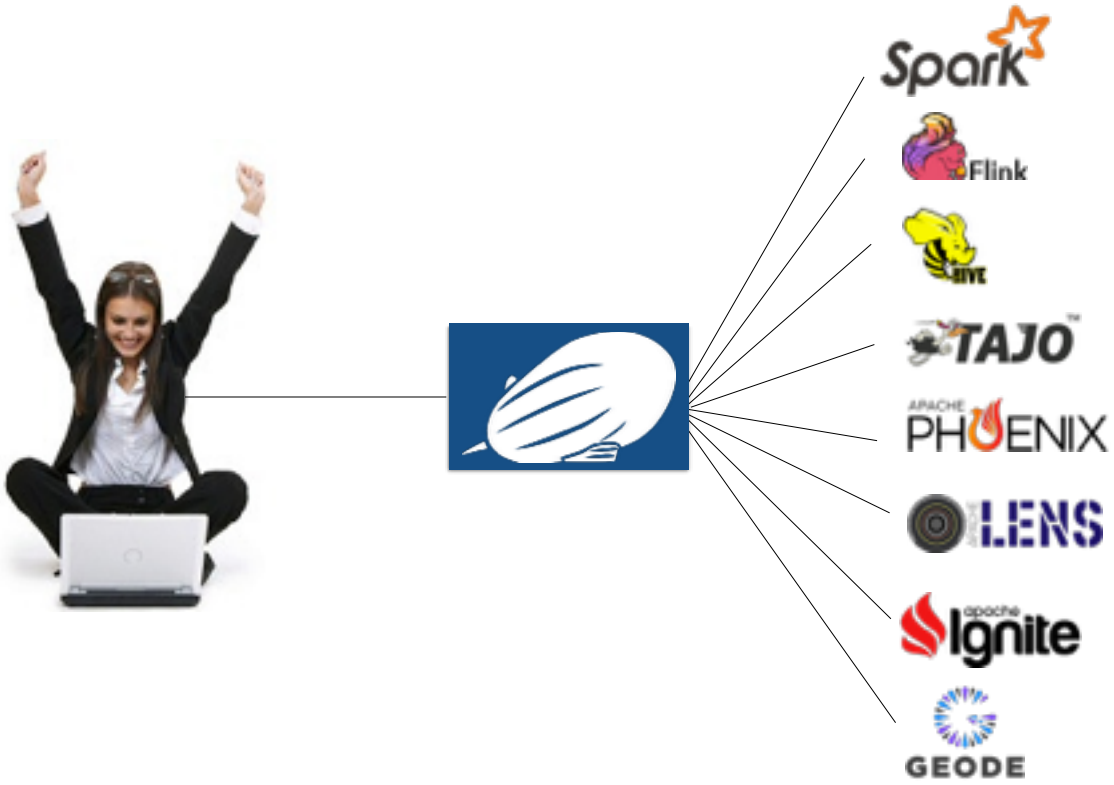
...

Ignite

LENS

TAJO

다양한 Backend 동시에 사용



다양한 Backend 동시에 사용

Shell 명령 이용해 데이터 카피

↓

Hive 이용해 데이터 transformation

↓

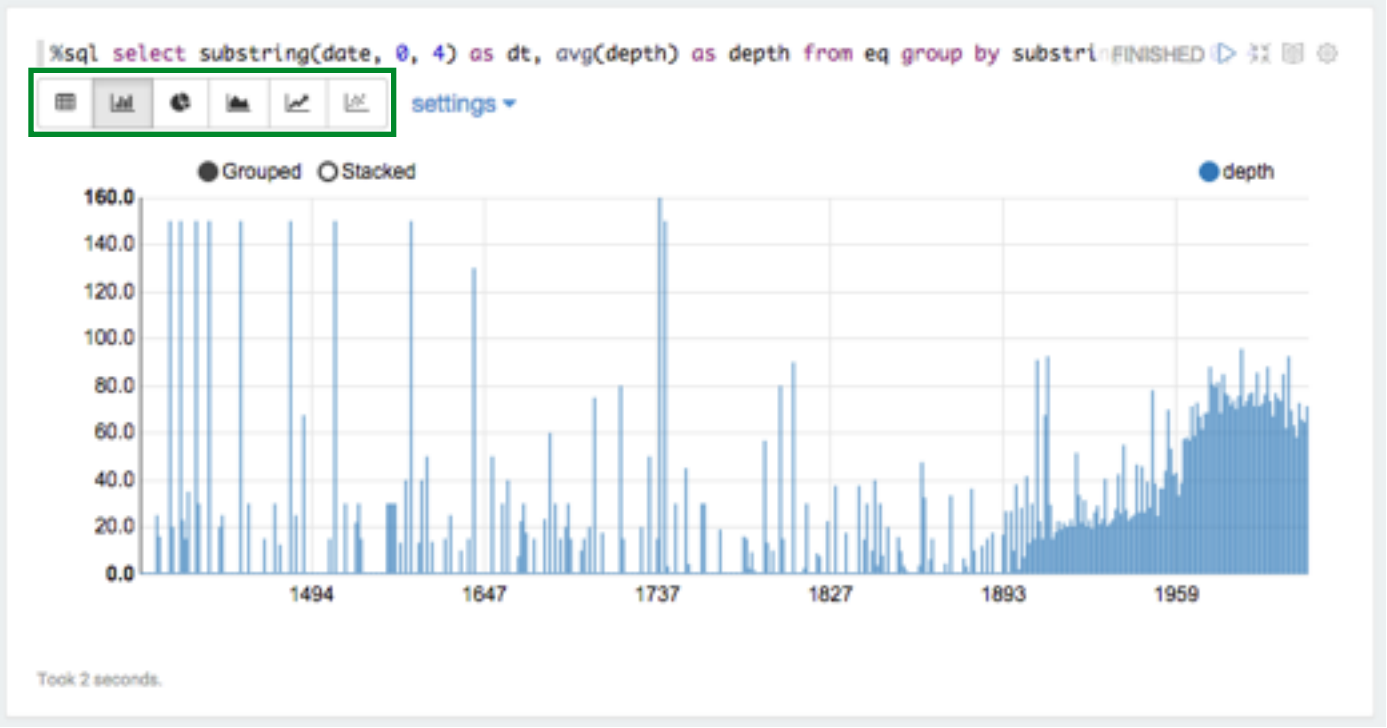
Spark 의 MLlib 으로 분석

↓

python 이용하여 시각화

하나의 노트북에서 순차적으로 작업을 처리

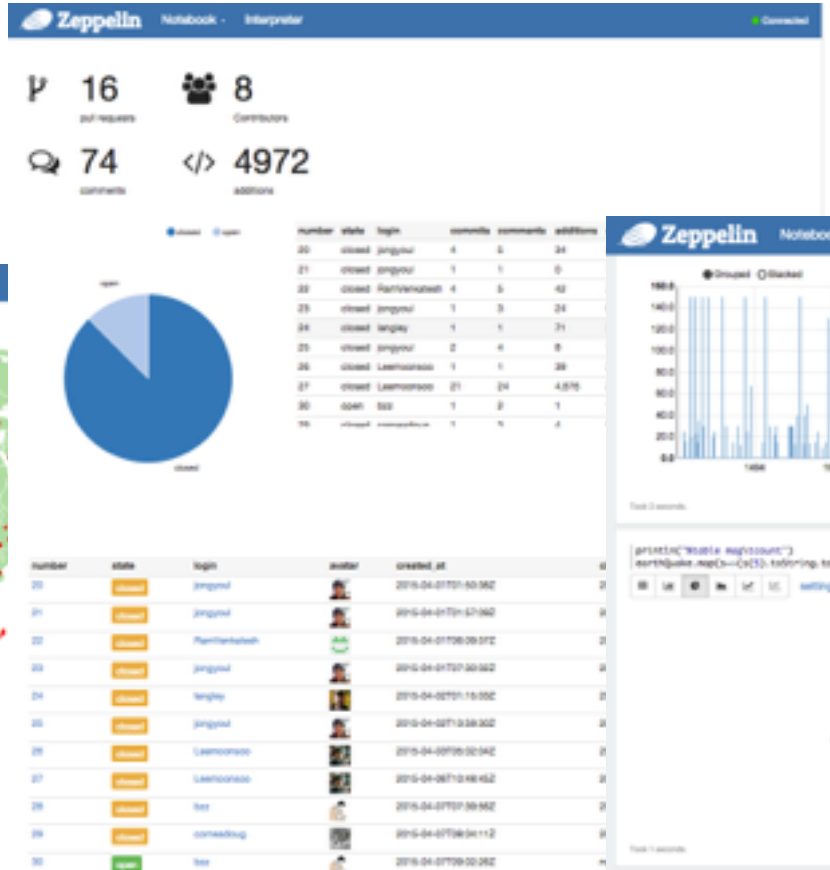
Visualization



Apache Zeppelin

DEVIEW 2015

Interactive Notebook

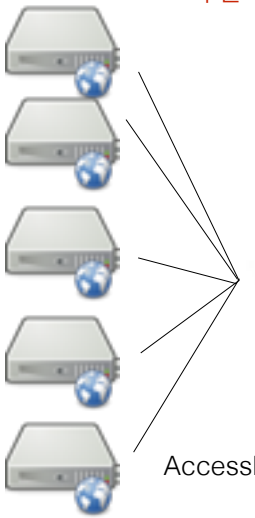


Before

Spark 에서 AdHoc query를 위해 Hive 에서 aggregation 을 미리 수행해야 하는 경우가 많은데, 소스가 따로 떨어져있어서 관리가 잘안되고 실수가 자주 발생

Excel 로 리포트 만들어 이메일 보내면 매일 매일 업데이트 해달라고 연락옴

데이터 넣는 과정에서 오류 발생하면 수동으로 재처리



AccessLog

30 대

400 GB / day

AdHoc

Batch

ETL / aggregation

PHP web server
High chart

Data parsing

Traffic normalize

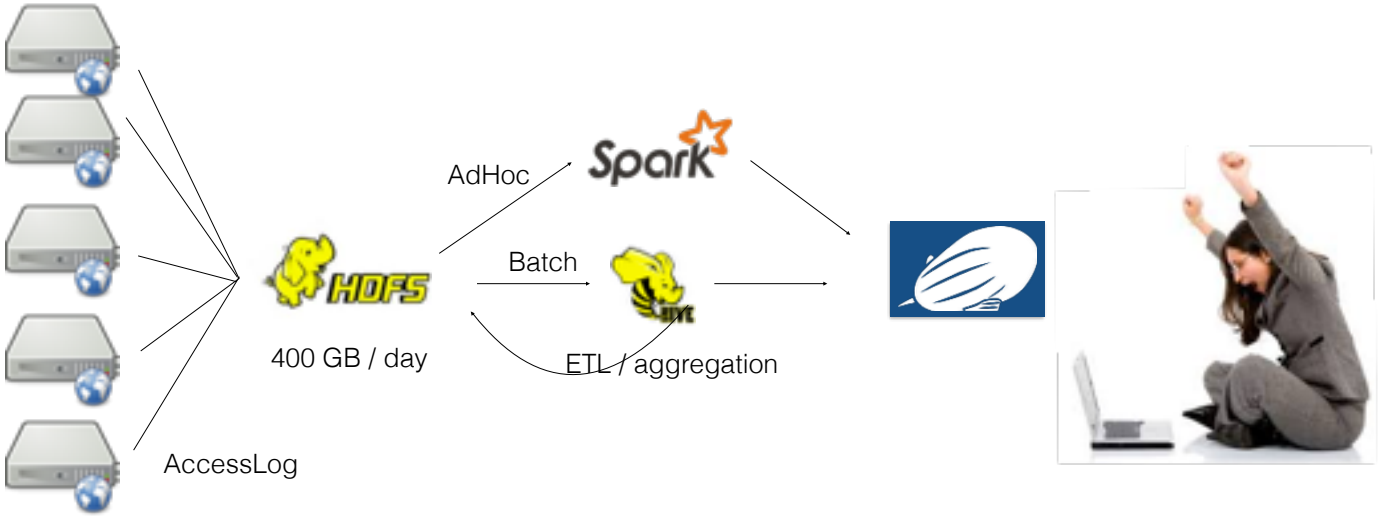
IP location lookup

Service account recognition

....

새로운 분석 추가될때 마다 mongodb 에 스키마 관리 웹페이지 수정 및 차트 추가

After


AdHoc
Spark
HDFS
Batch
400 GB / day
ETL / aggregation
AccessLog

.

.

.

.

3.
Zeppelin
in your team

Team = Multi user

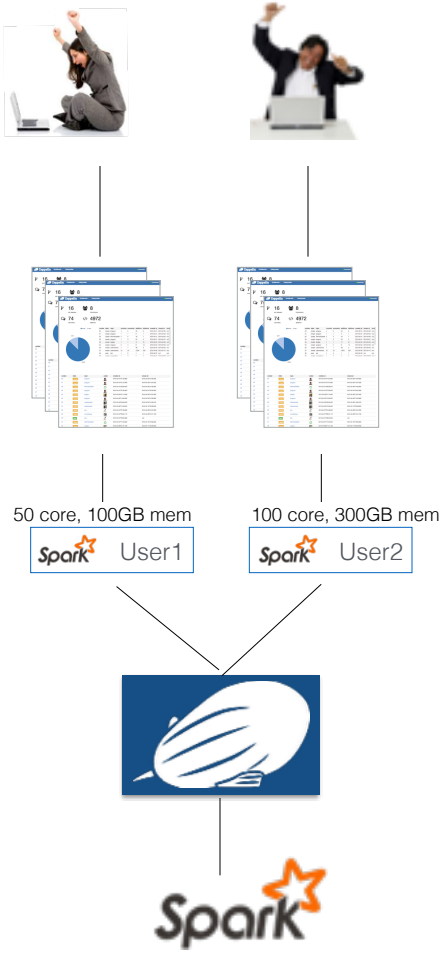
Multi user = ACL

다중 사용자 구성 A

각각 노트북에서 interpreter setting 선택 하여 사용

사용자별로 Interpreter setting 추가 하여 각각 다른 리소스 할당

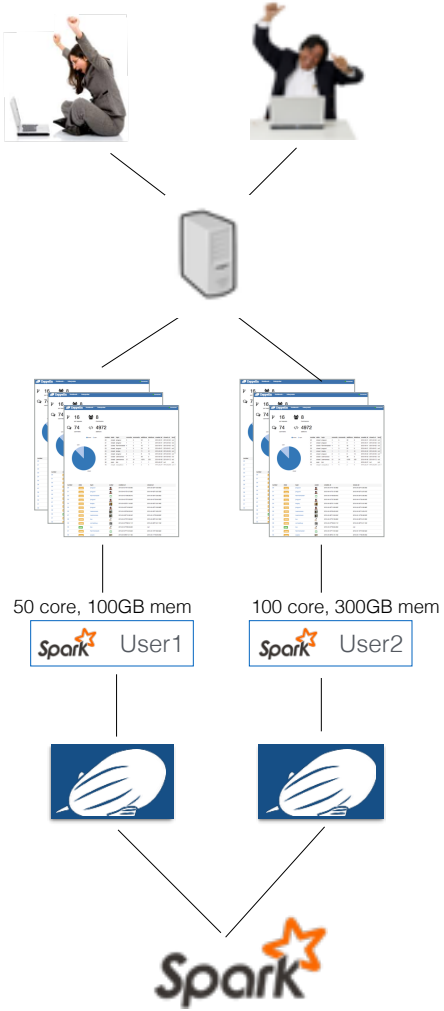
하나의 Zeppelin 인스턴스 사용

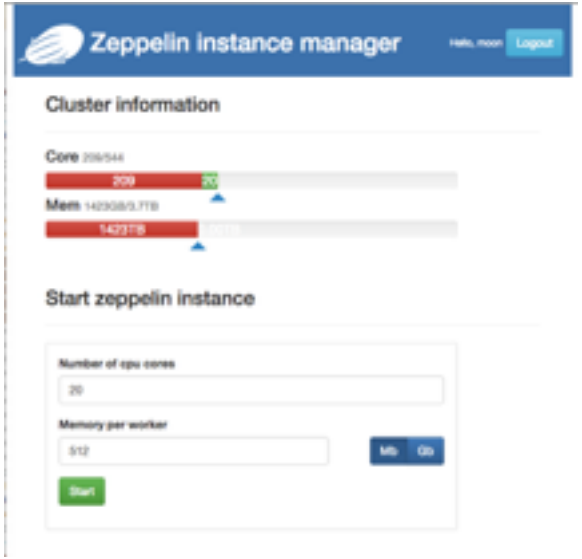


다중 사용자 구성 B

Proxy server 에서 인증 및 Instance 할당 관리

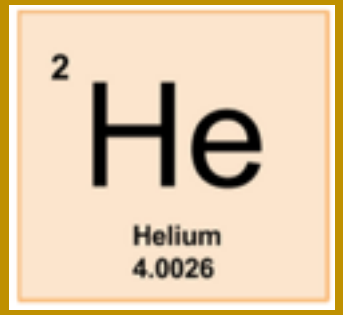
사용자별로 Instance 생성





http://nflabs.github.io/z-manager/

4.

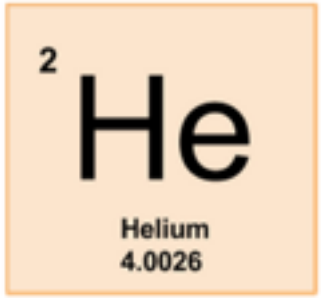


Platform for **Data analytics application**
on top of Apache Zeppelin

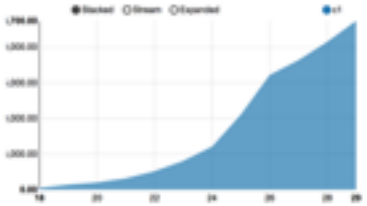
https://cwiki.apache.org/confluence/display/ZEPPELIN/Helium+proposal

Helium Application

DEVIEW 2015



=



+

View

Algorithm

Zeppelin provided Resources

Resources

Example of resource

Data
- Result of last execution
- JDBC connection (from JDBC Interpreter)*

Computing
- SparkContext (from SparkInterpreter)
- Flink environment (from FlinkInterpreter)*

Any java object
- Provided by user created Interpreter
- Provided by user created Helium application

예

데이터

- ex) get git commit log data
  https://github.com/Leemoonsoo/zeppelin-gitcommitdata

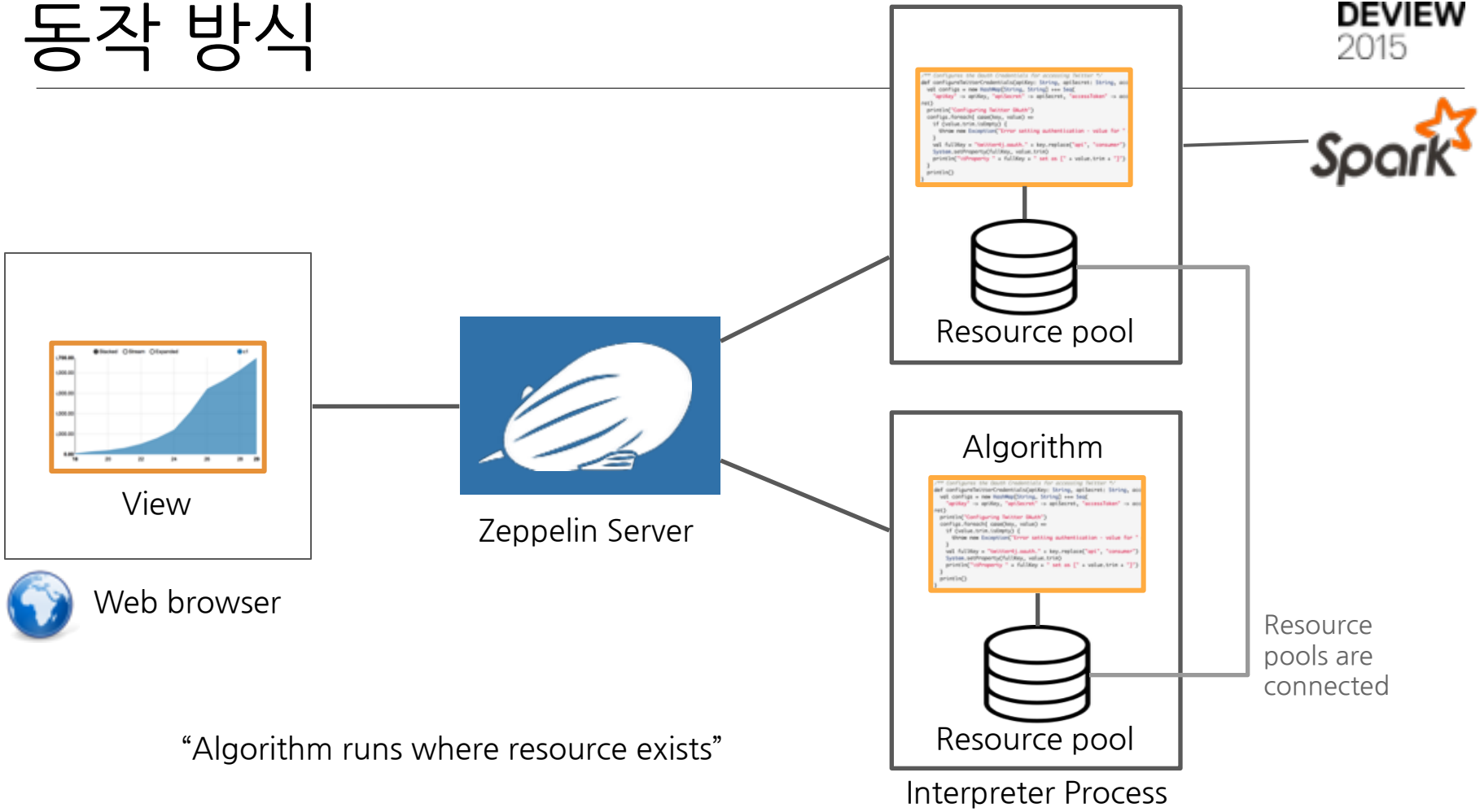
컴퓨팅

- ex) run cpu usage monitoring code across spark cluster, using SparkContext
  https://github.com/Leemoonsoo/zeppelin-sparkmon

시각화

- ex) display result data as a wordcloud
  https://github.com/Leemoonsoo/zeppelin-wordcloud

동작 방식


Spark
Resource pool
View
Zeppelin Server
Web browser
Algorithm
Resource pool
Interpreter Process
Resource pools are connected
“Algorithm runs where resource exists”

API

쉽다

```
class YourApplication extends org.apache.zeppelin.helium.Application {

  @Override
  public void run(ApplicationArgument arg, InterpreterContext context) {
    .....
  }
}
```

하나의 class 만 extend 하면 만들수 있다.

Application Spec

DEVIEW 2015

간단하다

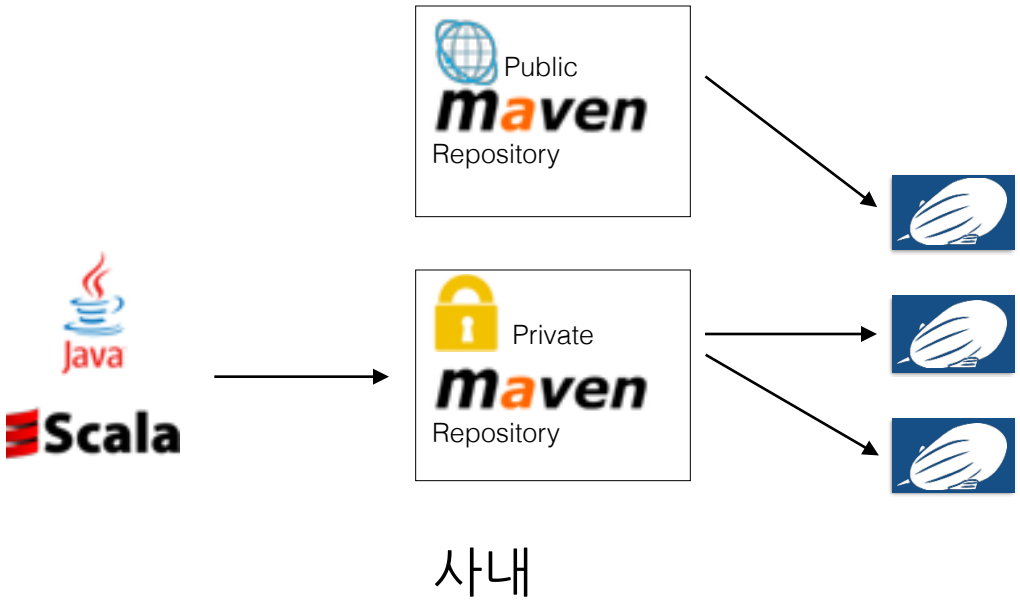
```
{
  mavenArtifact : "groupId:artifactId:version",
  className : "your.helium.application.Class",
  icon : "fa fa-cloud",
  name : "My app name",
  description : “some description",
  consume : [
    "org.apache.spark.SparkContext"
  ]
}
```

Spec 파일만 정의하면 바로 사용

Deploy

편리하다

공개



maven 사용해서 배포되고

설치되므로 사내 라이브러

리를 따로 구축할 수 있다.

사용자가 클릭하면 바로 다

운로드되어 실행

사내

maven 통해서 배포하여 편리하게 사용

Demo

Get involved

홈페이지
http://zeppelin.incubator.apache.org/

메일링리스트
users@zeppelin.incubator.apache.org
dev@zeppelin.incubator.apache.org

이슈트래커
https://issues.apache.org/jira/browse/ZEPPELIN

소스리파지토리
https://github.com/apache/incubator-zeppelin

Q&A

Thank You