



powered by **Spark**

Apache Zeppelin

으로 데이터 분석하기

2015-01-19

스사모 (한국 스파크 사용자 모임)

<https://www.facebook.com/groups/sparkkoreauser/>

김상우, VCNC(비트윈), Zeppelin 커미터
kevin@between.us, kevinkim@apache.org

Zeppelin

- Early stage 프로젝트 (Github 50 Star)
- 1~2년 사이에 엄청 유명해질 프로젝트
- 10줄만 커밋해도 contributor 로 넣어주는 좋은 프로젝트
- 쉬운 설치, 실행하면 Spark을 내부에서 띄워줌 (외부 Cluster와 연결도 가능)

3개월 전...

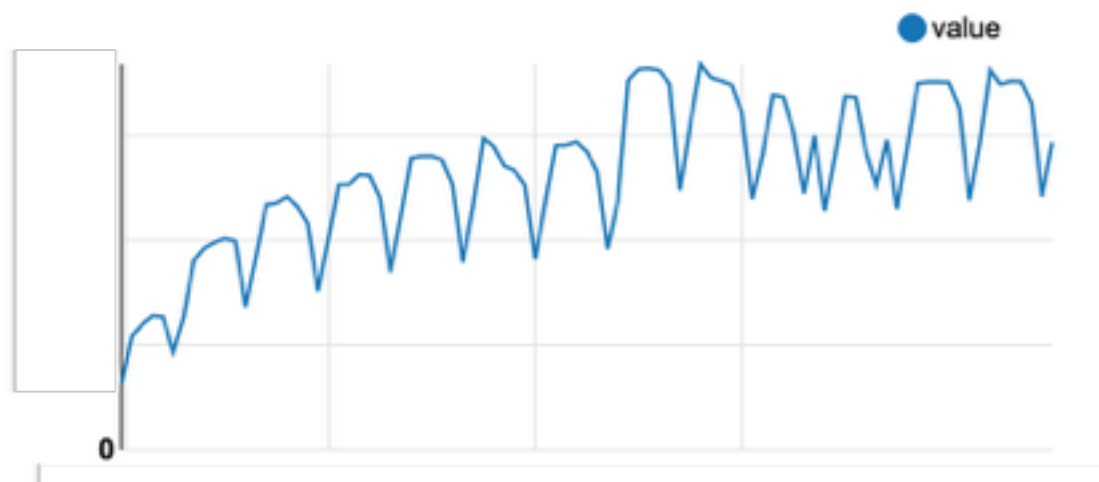
Apache Zeppelin

- 데이터 분석가, 개발자들을 위한 웹기반 노트북, 시각화 툴
- Spark, SparkSQL의 결과를 바로 차트로 그릴 수 있음
- 2014년 12월에 Apache 소프트웨어 재단의 Incubating 프로젝트가 됨
- Apache Tajo, Apache Flink 등 다양한 엔진을 결합 시도

FINISHED ▶ ⌵ ⌵ ⌵ ⌵

```
%sql
select date, value
from daily_stats_summary
where date >= "2014-10-20"
and action = 
order by date
```

				
---	---	---	---	---

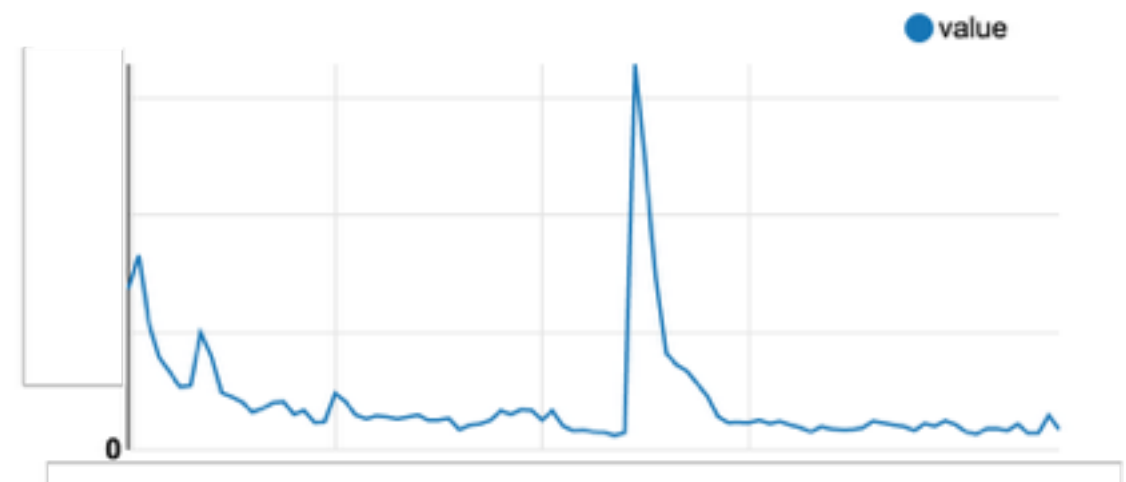
SETTINGS ▾


Took 0 seconds

FINISHED ▶ ⌵ ⌵ ⌵ ⌵

```
%sql
select date, sum(value) value
from 
where date <= todayDateDailyStatsSummary(0)
group by date
order by date
```

				
---	---	---	---	---

SETTINGS ▾


Took 1 seconds

기존의 Workflow



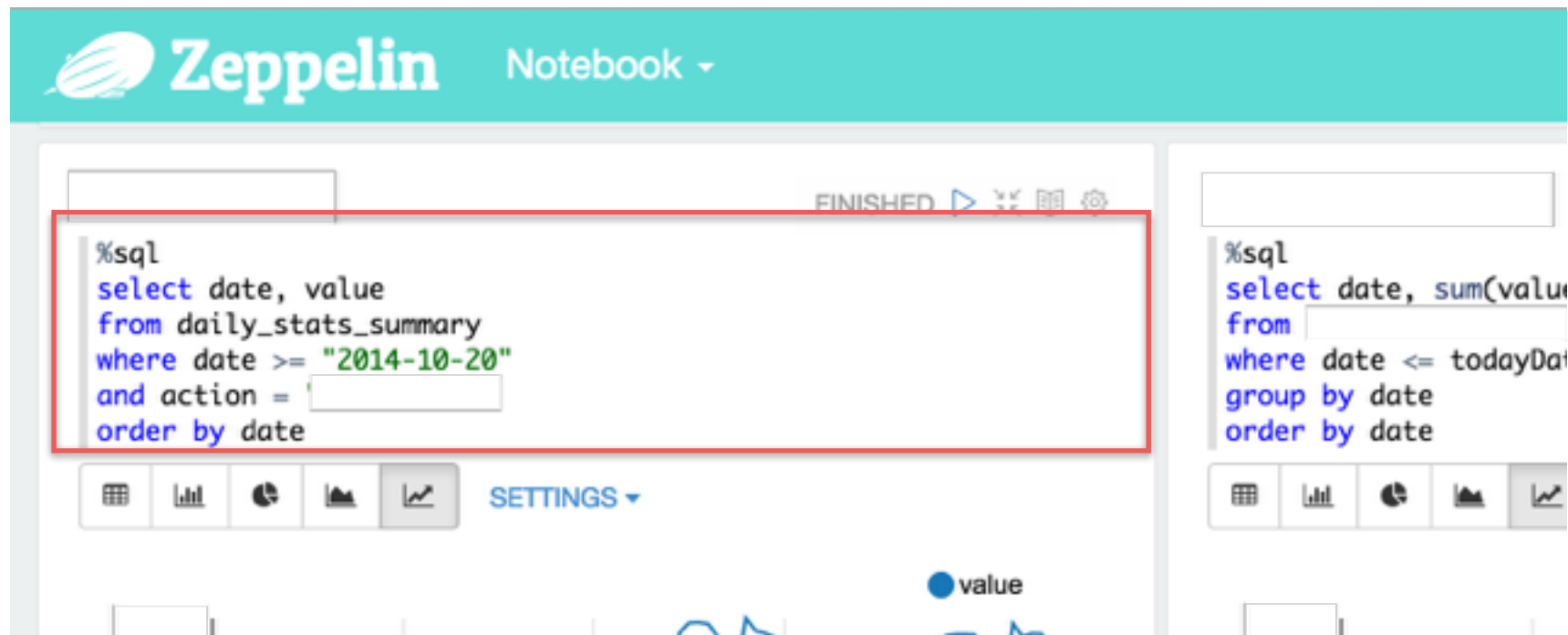
- 다양한 제품을 조합하여 데이터 분석을 하는것이 일반적
- 많은 엔지니어링이 필요함
- 다방면에 경험 많은 분석가들 혹은 팀의 전유물
- 파이프라인이 복잡하기에, 고장나기 쉽고 유지보수 어려움

새로운 Workflow



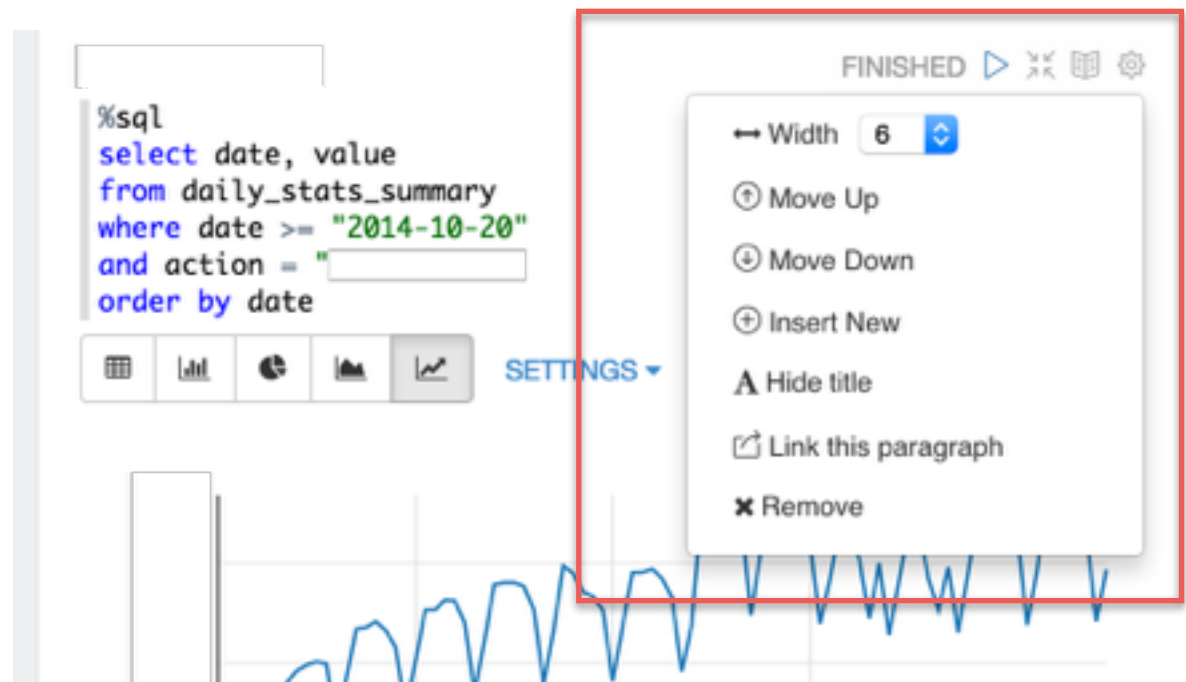
데이터 정제, 처리, 요약 데이터 시각화,
고급 분석까지 전부 Spark와 Zeppelin으로 해결

Notebook



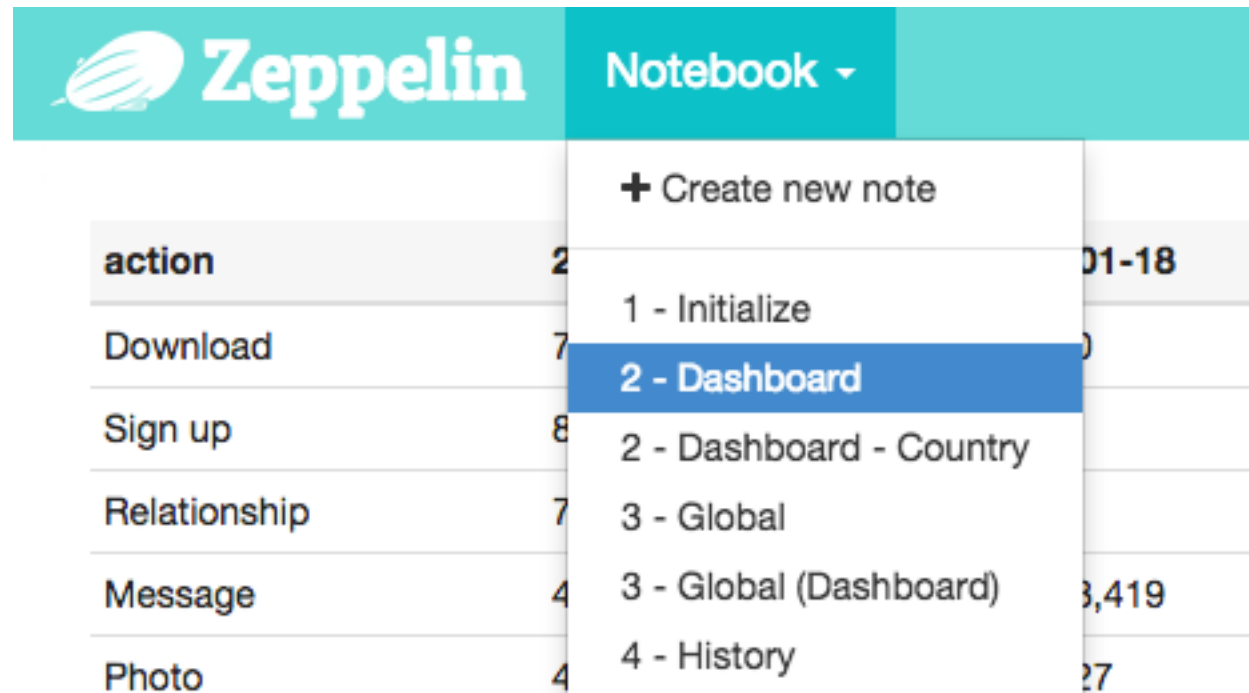
- 소스코드 작성, 수정, 자동저장, 실행
- Scala (Spark), Spark SQL, Markdown 등 지원

Notebook (2)



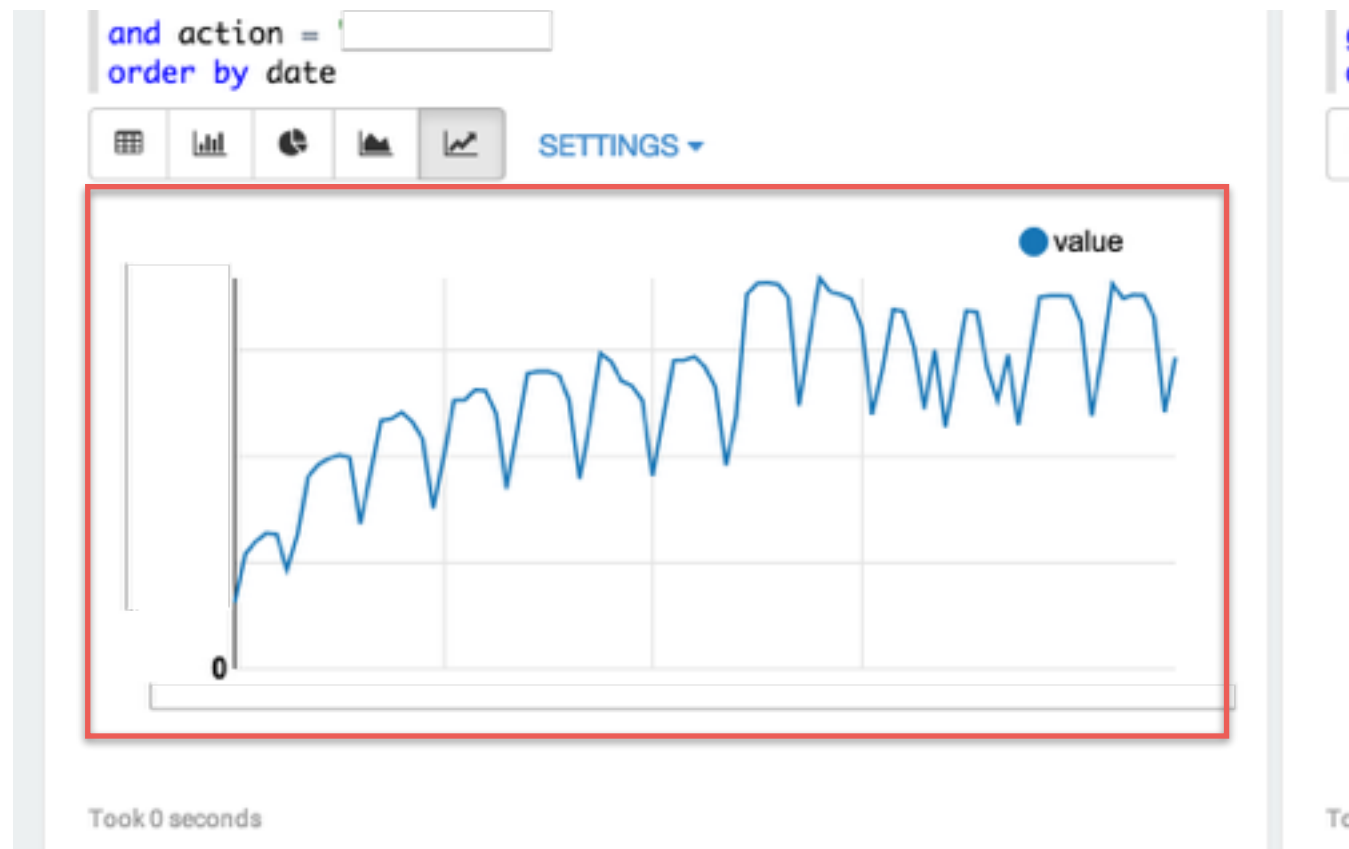
- Paragraph들의 실행 상태를 컨트롤
- Paragraph들의 모양 및 위치 조정, 제목 표시 등 편집 가능

Notebook (3)









- 여러개의 노트북을 생성, 목록으로 관리 가능
- 분석 작업 코드 및 결과물을 효율적으로 관리

Visualization



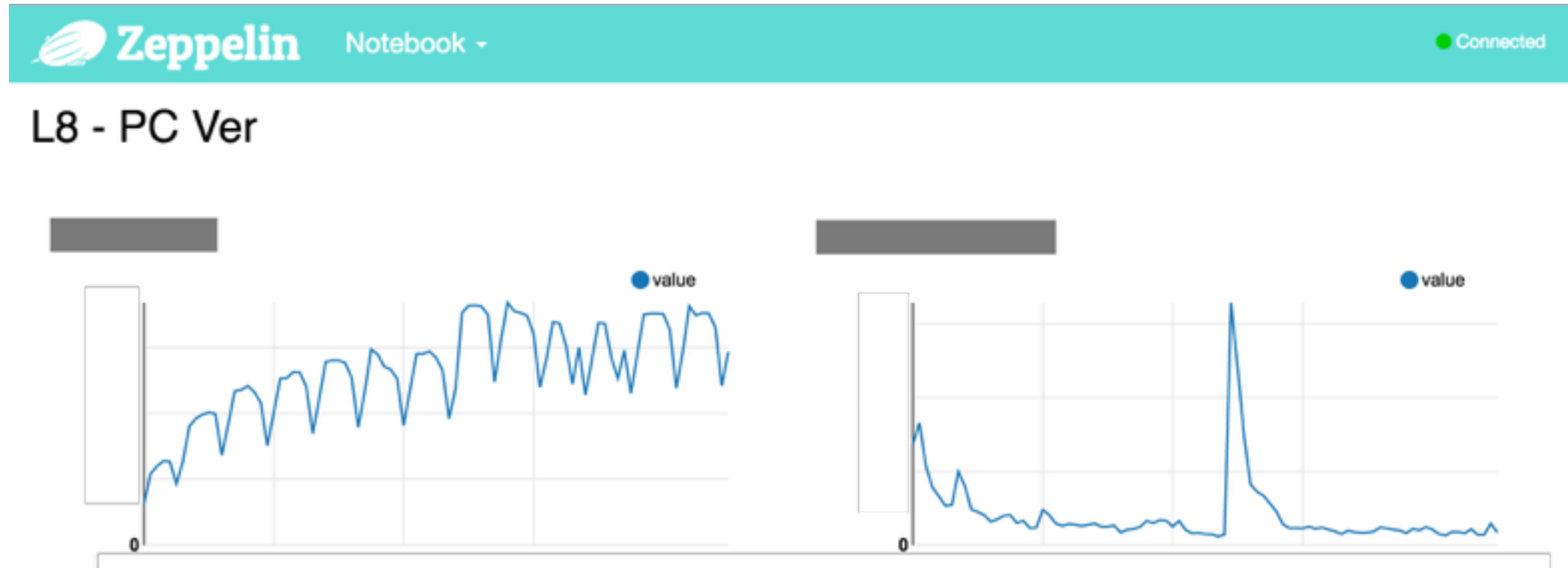
- Spark SQL 수행 결과를 Table, Line Chart, Pie Chart 등 다양한 형태로 시각화
- Spark의 좋은 성능 덕분에 대부분 코드가 즉시 실행되므로 interactive 하게 데이터를 다룰 수 있게 됨

Visualization (2)

date	image	value	set_id	price	name
2015-01-19			82	2.99	Merry&Milk
2015-01-19			106	2.99	Robin Egg Special Ed..
2015-01-19			101	2.99	Mochi Special Editio..
2015-01-19			59	2.99	Peng & Mr. White, Se..
2015-01-19			48	1.99	Merry,Milk,Ivy and G..

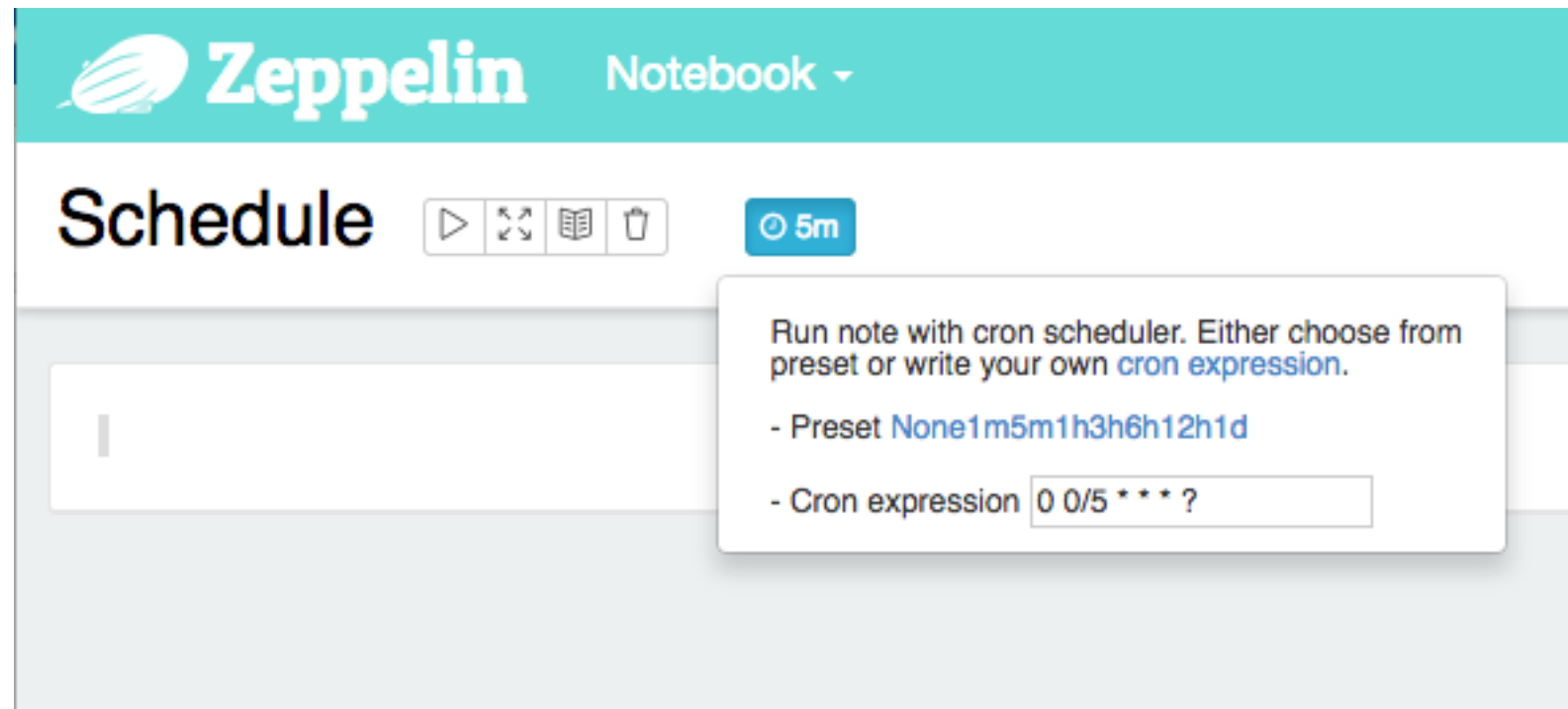
- HTML을 표현 가능하므로, 테이블에 이미지를 표시하거나, link를 넣거나 하는 등의 동작이 가능
- SparkSQL의 간편한 UDF(User Defined Function) 등록 기능과 결합하면 편리함

Dashboard



- Default, Simple, Report 뷰 모드를 제공함
- 코드를 가려주는 Report 뷰 모드를 활용하면 Dashboard를 빠르게 만들수 있음
- 코드와 차트들이 한군데 있으므로 손쉽게 페이지를 새로 만들고, 유지 관리 가능

Dashboard (2)



- 자체적으로 Schedule 기능 내장
- 매일 혹은 매 시간 업데이트 되는 Dashboard나, Batch작업을 관리하기 용이함

Live Demo

Zeppelin을 추천합니다

- 간단하게 데이터 분석을 시작해보려는 사람
- Spark을 처음 시작하려는 사람
- Dashboard를 빠르게 만들고 싶은 사람
- 민첩하게 이런저런 데이터를 살펴보고 분석하는 작업
- 오픈소스 프로젝트에 참여해보고 싶은 사람

감사합니다