KossLab
KOREA OPEN SOURCE SOFTWARE DEVELOPERS LAB
공개SW개발자 Lab

빅데이터 처리툴 Spark 의 실행과
데이터 시각화를 지원하는 툴

Apache Zeppelin

공개 SW 개발자 Lab 오픈소스프론티어 3 기 윤제상

KossLab

최근 빅 데이터의 중요성과 데이터 분석의 필요성이 널리 알려지면서 분석 기법뿐만 아니라 사용되는 도구(툴/시스템/프레임워크) 등이 주목을 받고 있다. 그 중에 단연 가장 많이 화두가 되는 도구는 Hadoop과 Spark일 것이다. 이들을 통해 과거에는 엄두도 내기 힘들었던 막대한 데이터 처리 및 분석을 오픈소스와 컴퓨팅 자원 그리고 어느 정도의 SW 개발에 대한 지식만 있으면 쉽게 할 수 있게 되었다.

그러나 이들 도구를 좀 더 잘 다루려면 일정 수준 이상의 SW 개발에 대한 경험과 지식이 있어야만 사용 가능하다는 단점이 있어 분석가가 좀 더 직관적으로 다룰 수 있는 도구가 필요하게 되었다. 분석 코드 작성과 실행 그리고 문서화를 한 곳에서 가능케 해주는 Notebook 스타일의 분석 도구는 이러한 점에서 최적의 도구로 주목을 받고 있으며, Apache Zeppelin은 그 중에서 가장 Hot한 오픈소스 프로젝트로서 주목받고 있다.

이번 시간엔 이러한 Apache Zeppelin이 어떻게 만들어지게 되었는지, 어떻게 구성되어 있고, 어떻게 설치하고 어떤 기능을 사용할 수 있는지에 대해서 대략적으로 다뤄보도록 하겠다.

프로젝트명	Apache Zeppelin
개요	Hadoop, Spark 및 다양한 오픈소스 기반 Data 프레임워크/애플리케이션들을 서로 조합하여 Notebook 스타일로 한 곳에서 분석코드 작성/실행/시각화/공유를 가능케 해주는 빅데이터 분석 도구
특징	- Apache Software Foundation Top Level Project - Hadoop, Spark 외 MySQL, BigQuery, Cassandra 등의 다양한 분석 도구 및 Database 운용 - IPython, Jupyter 처럼 Web 기반 UI 에서 데이터 분석 코드 작성/실행/시각화/공유가 가능 - 여러 사람이 동시에 분석 코드 작성을 같이할 수 있는 협업 기능 존재 - Plugin 방식의 구조를 채택하여 자신의 입맛에 맞는 새로운 기능을 추가 개발 및 적용 가능 - Apache 계열 오픈소스에서 한국인이 주도하는 몇 안 되는 프로젝트 중 가장 Hot 한 프로젝트
목표	- 빅데이터 분석 도구 생태계의 발전: 기업 및 조직 내의 다양한 데이터 관리 및 분석 프레임워크들을 단일 애플리케이션(Zeppelin)에서 서로 융합하여 사용하며 조직 내에 데이터 분석 결과를 Web 을 통해 쉽게 공유할 수 있도록 함
기대효과	빅데이터 분석 도구 생태계의 주요 Hub 역할을 기대
리퍼지토리	https://github.com/apache/zeppelin

[목차]

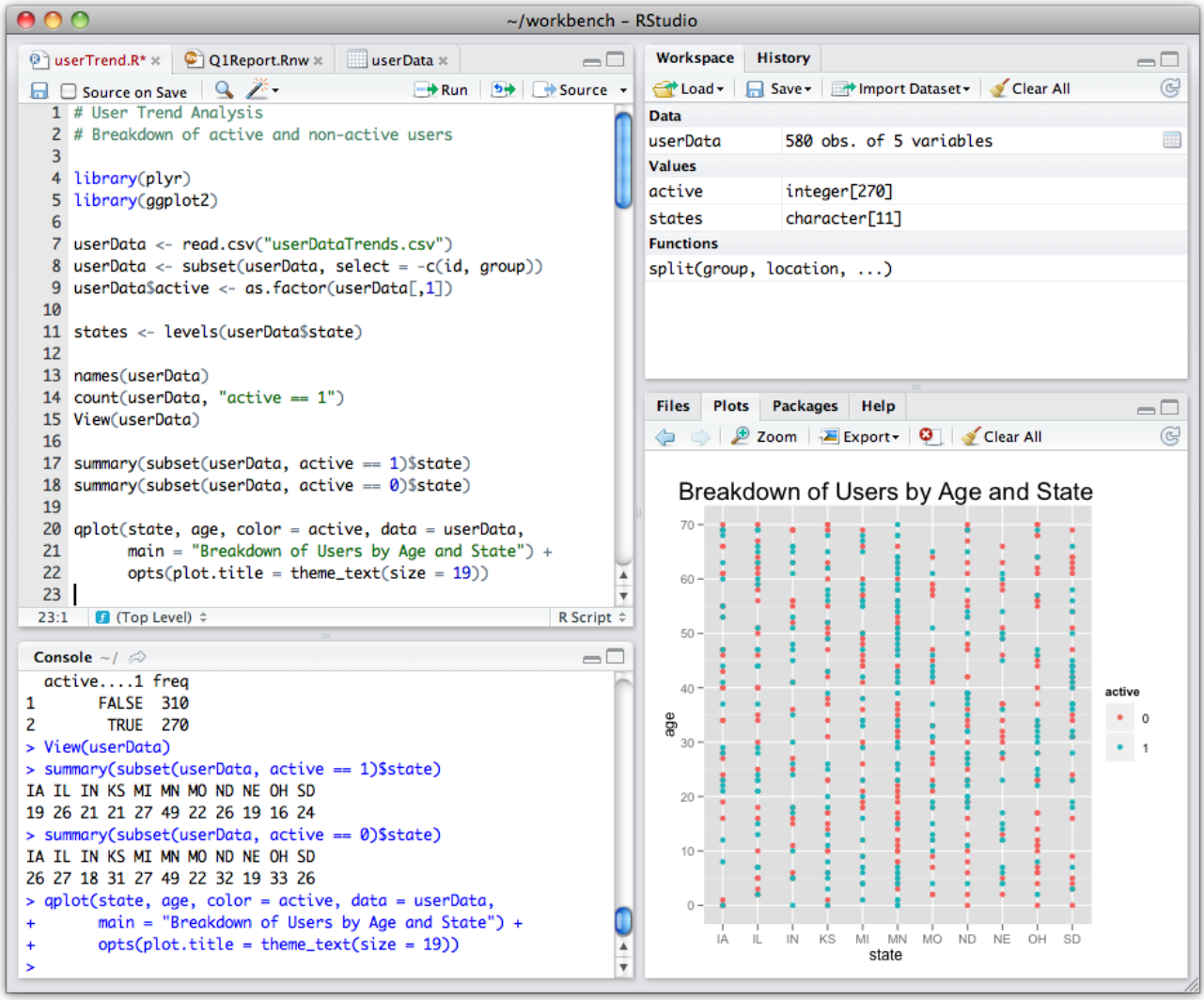
1 탄생 배경

1.1 Hadoop/Spark의 접근성 문제

Hadoop에 이어 가장 핫한 빅데이터 분석 도구인 Spark는 Hadoop보다 빠르며 설치가 비교적 쉽다는 장점을 가지지만 이를 다루려면 Hadoop과 마찬가지로 커맨드라인 콘솔에서 작업해야 한다. 하지만 데이터 분석 업무를 하다 보면, 여러 가지 알고리즘과 파라미터를 바꿔가면서 분석 결과를 그래프나 테이블 같은 형태로 시각화하여 확인하는 작업을 수도 없이 반복하게 되는데 커맨드라인 콘솔은 불편하고 작업의 속도를 떨어뜨릴 수 있다. 비슷한 예로 데이터 분석 분야에서 표준이 된 R 언어 역시 CRAN 패키지만 깔면 커맨드라인으로 모든 걸 할 수 있으나 실무에선 분석 결과를 GUI를 통해서 확인하고 Graph 같은 요소들을 쉽게 볼 수 있도록 도와주는 RStudio등이 널리 쓰이고 있다.

[그림 1] Spark만 설치했다면 이런 인터페이스에서 분석 코드를 짜야 한다.

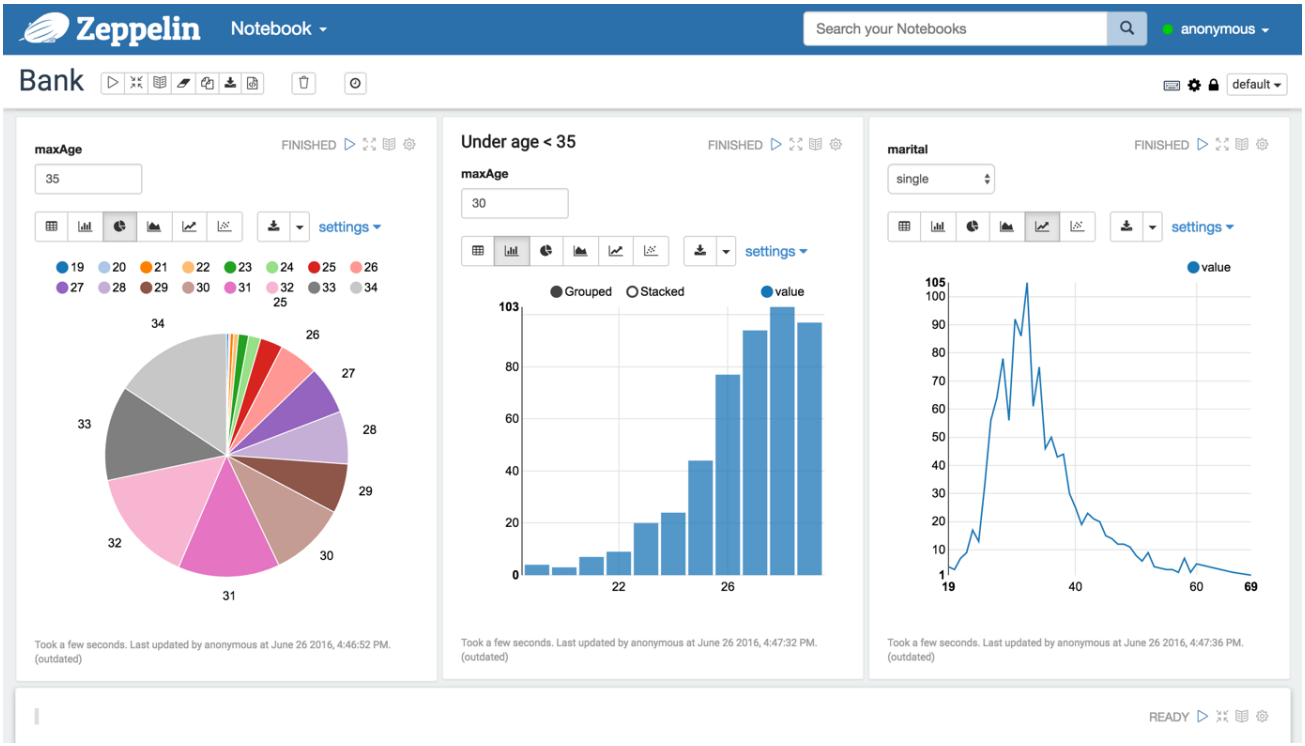
(출처: http://yokekeong.com)



[그림 2] RStudio는 R을 매우 편리하게 사용할 수 있도록 도와준다.

(출처: http://www.rstudio.com)

Apache Zeppelin은 Spark를 통한 데이터 분석의 불편함을 Web 기반의 Notebook을 통해서 해결해보고자 만들어진 애플리케이션이다. Web 기반 Notebook 환경이란 Web 브라우저에서 워드프로세서처럼 아무거나 입력 가능한 화면에 코드를 작성-실행-결과확인-코드수정을 반복하면서 원하는 결과를 만들어 낼 수 있는 작업환경을 말한다.



[그림 3] CLI 마니아가 아니고서야 이런 GUI에 끌릴 수밖에 없을 것이다.

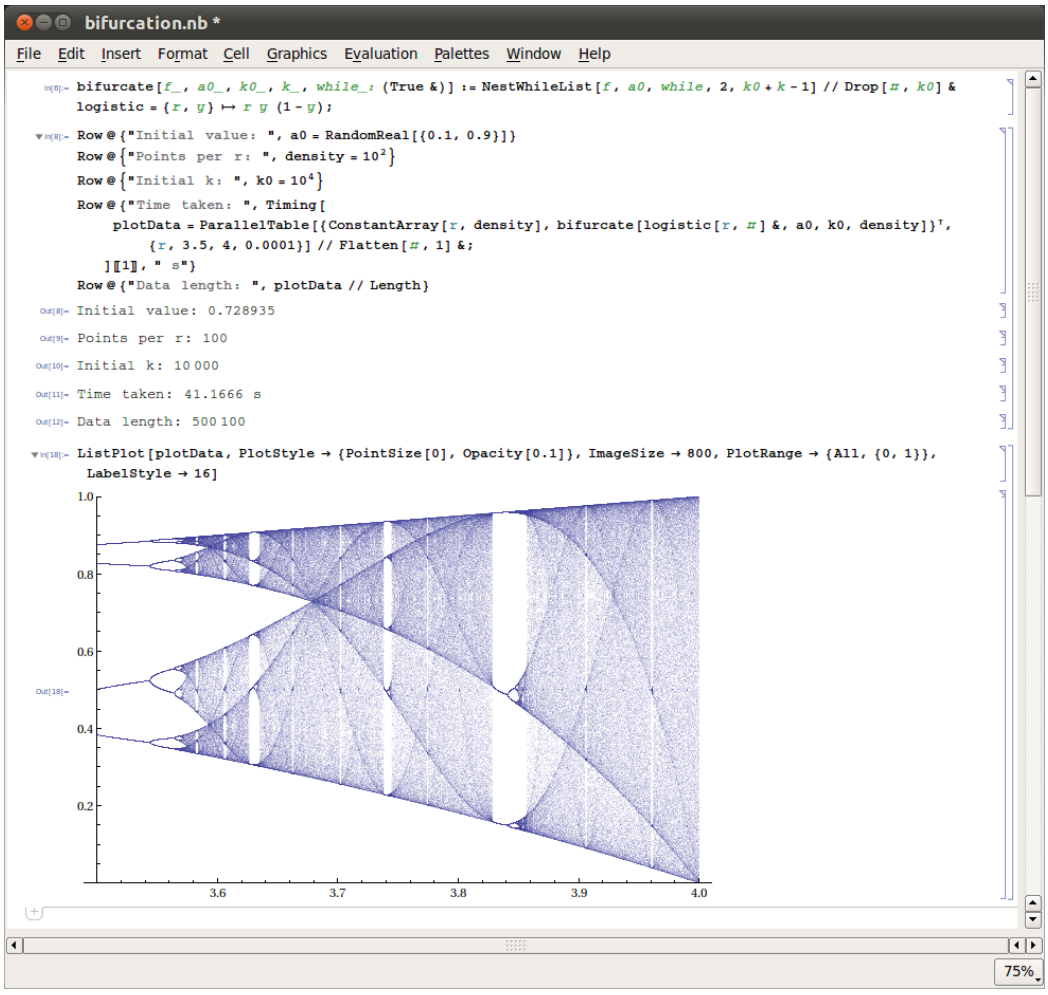
(출처: http://zeppelin.apache.org)

1.2 Notebook 환경을 가진 분석 도구의 역사

사실 Notebook 형태로 데이터를 다루고 분석하는 애플리케이션은 Zeppelin이 최초는 아니다. 20년 전부터 Wolfram Research의 Mathematica나 National Instrument사의 Matlab이 이러한 Notebook 환경을 선도해 왔다. Notebook 환경은 매우 큰 자유도를 제공하기 때문에 Mathematica나 Matlab의 유저들은 코드작성뿐만 아니라 일반문서 및 보고서, 데모를 위한 프레젠테이션까지 이 환경 안에서 구현하여 쓰고 있을 정도다. 잘만 사용하면 어떤 GUI 인터페이스보다 강력한 일들을 할 수 있는 것이 Notebook 환경이다.

[그림 4] Wolfram Research의 Mathematica는 1988년부터 Notebook 환경을 제공해 왔다.

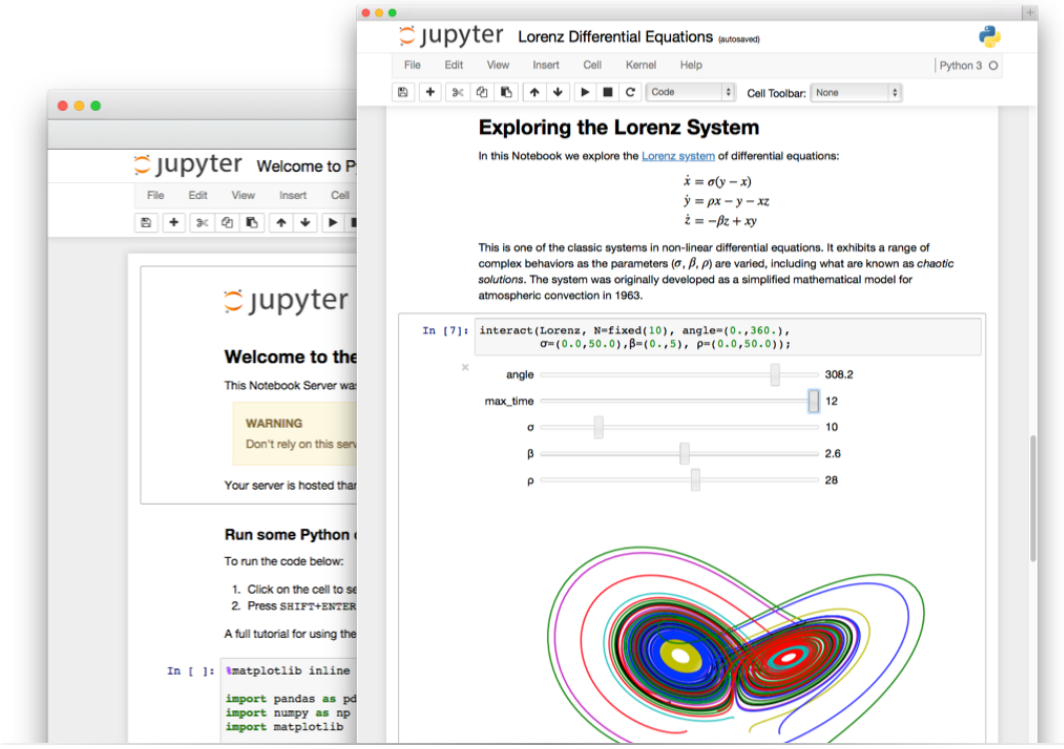
(출처: http://www.mathematica25.com/)



[그림 5] Mathematica의 Notebook 환경에선 데이터와 분석코드 그리고 그래프를 자유롭게 섞어 쓰고 배치할 수 있다. (출처: 위키피디아)

다만 이런 환경은 너무 큰 자유도 때문에 초심자가 직관적으로 사용하기가 어렵다는 문제도 가지고 있다. 처음 접하는 사용자들은 빈 공책에 뭘 채워 넣을지 고민해야 하는 것처럼 빈 화면에 뭘 입력해서 시작해야 할지 고민하게 된다. 그런 상황에서 원하는 결과를 뽑을 때까지 끈기있게 예제를 탐독하고 실패를 여러 번 경험하고 난 후에야 어떻게 오답 노트를 공책에 잘 정리할 수 있는지 알게 되는 것처럼 쓸 수 있게 된다. 이 때문에 이런 환경을 자유롭게 쓰는 유저는 연구소나 대학 또는 금융권의 전문가(이라 쓰고 하드코어 유저라 읽는다)들이 대부분을 차지하고 있다. Mathematica의 개발팀이었던 필자 지인의 경험에 따르면 Wolfram Research 내에서도 비전문가인 사용자를 늘리기 위해 Notebook 인터페이스를 포기해야 하느냐에 대해서 오랜 논의가 있었다고 한다.

이뿐만 아니라 이러한 애플리케이션들이 오랜 시간 동안 오픈소스 형태의 무료로 제공되기보다, 비싼 돈을 내고 써야 하는 기업의 제품으로 제공되어 왔기 때문에 기업 외의 일반 유저들이 접하기는 쉽지 않았다. 그래서 그런지 Notebook 형태의 분석 툴은 그 강력한 기능과 PC의 여명기에 등장했음에도 불구하고 오랜 시간 동안 대중화되기 어려웠다. 그러다가 최근 들어 빅데이터 붐이 불고, Hadoop을 시작으로 빅데이터 분석 툴이 주목을 받기 시작하면서 이를 사용하는 유저층이 넓어졌고 커맨드라인 대신에 편리하게 분석 툴을 사용할 수 있는 방법을 찾기 시작했다. 빅데이터 분석의 주 수요층인 연구소나 대학의 데이터 분석가들에게 익숙한 인터페이스인 Notebook 환경이 주목받기 시작하면서 오픈소스 진영에서 IPython, Jupyter, Zeppelin 등이 등장했다.



[그림 6] Jupyter는 Python, Julia 등의 다양한 프로그래밍 언어로 Notebook을 작성할 수 있게 도와준다. (출처: http://Jupyter.org)

2 프로젝트 구조

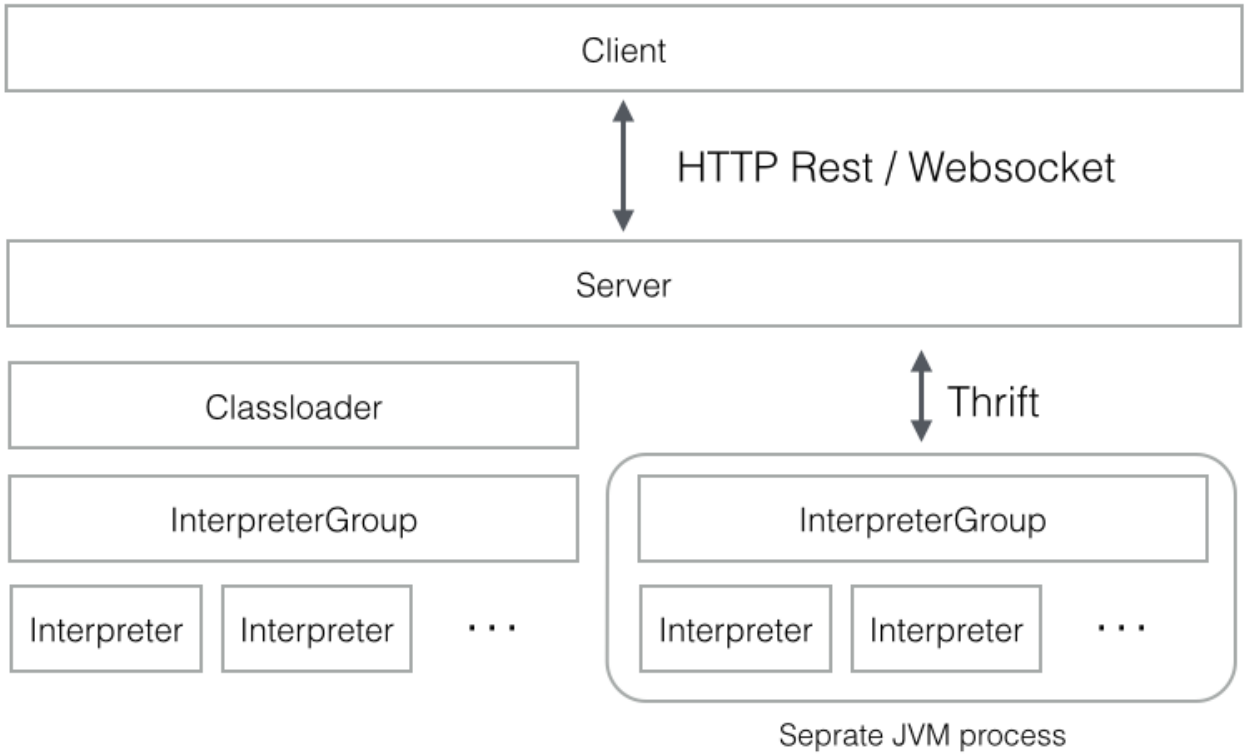
2.1 Zeppelin 과 데이터 분석 도구 생태계

[그림 7] Zeppelin에서 Interpreter를 통해 다룰 수 있는 다양한 기술 스택들

(출처: http://zeppelin.apache.org)

Zeppelin은 비교적 최근에 등장한 분석 도구로서 아파치 소프트웨어 재단의 후원을 받으며 개발이 되고 있다. 사용자들은 Zeppelin을 이용하여 Web에서 Python, Scala 등의 다양한 언어를 섞어가며 분석 코드를 짤 수 있고 이 결과를 바로 Graph로 시각화하여 볼 수 있다. 여기에 더해 Zeppelin은 넓어져 가는 빅데이터 분석 도구 시장의 다양한 니즈에 맞추기 위해서 Spark뿐만이 아닌 Livy, Cassandra, Lens, SQL 등등의 다른 데이터 분석 도구나 데이터베이스에 접근하여 쿼리하는 것을 쉽게 할 수 있는 확장 기능들을 지원한다. 오픈소스를 기반으로 빅데이터 분석 시스템을 구성하는 기업들은 다양한 기술 스택을 서로 엮어서 시스템을 구성하게 되는데 Zeppelin 하나만 있으면 이들 시스템의 각 요소에 자유롭게 접근하여 데이터를 다룰 수 있게 된 것이다. 이러한 확장성 때문에 Zeppelin은 후발주자임에도 매우 빠르게 여러 기업에서 사용되기 시작했고 널리 알려지게 된 것이다.

2.2 Zeppelin 의 구조



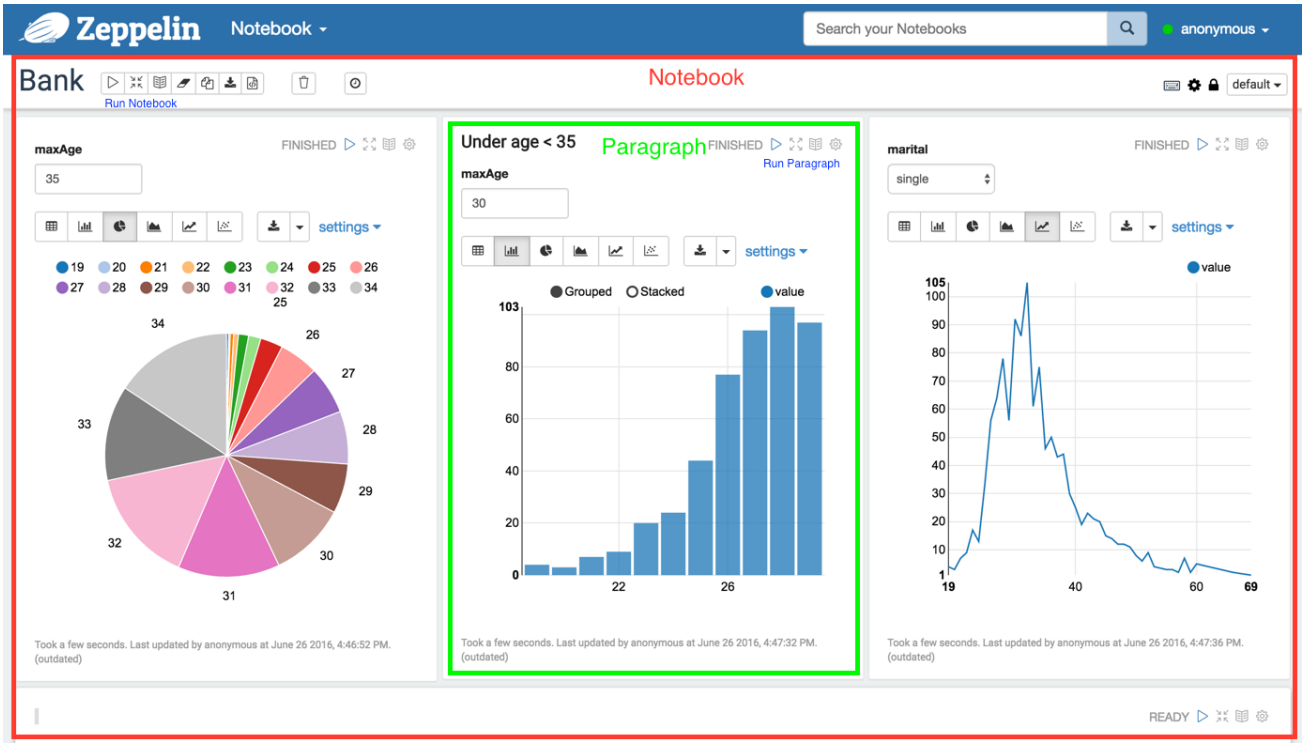
[그림 8] 인터프리터를 통해서 서로 다른 기술 스택도 같은 Web에서 다룰 수 있도록 해준다.

(출처: http://zeppelin.apache.org)

이러한 확장성은 Zeppelin의 Interpreter라는 플러그인 구조로 지원되는데 각 Interpreter는 Zeppelin의 Web Interface를 통해서 입력받은 분석 코드를 local 또는 원격에서 실행할 수 있다. 예를 들어 Spark로 Map-Reduce하는 코드를 작성하고 실행을 누르면 Zeppelin 안에 설치된 Spark Interpreter가 이를 받아 Spark Master에 Client 라이브러리를 통해 코드를 보내고 그 실행 결과를 받아 다시 Web Interface에 보내준다. 또한 Bash로 쉘 스크립트를 짜면 Zeppelin 안에 탑재된 Shell Interpreter가 이를 받아 Zeppelin이 설치된 서버에서 shell script를 실행하고 그 결과를 Web Interface에 보내주는 형태이다. Zeppelin 자체가 데이터 분석 처리를 하지 않기 때문에 분석 시스템이나 데이터베이스 등이 미리 구성되어 있어야 하고, Zeppelin과 이들 시스템을 연결해주는 작업을 해야 한다는 어려움이 있지만 한번 연결해 두면 같은 Notebook에서 Pyspark와 SparkR을 쓰거나 Cassandra DB 등의 데이터베이스까지 다루는 일을 Zeppelin 내에서 편하게 할 수 있는 장점이 있다.

3 주요기능

3.1 분석코드 작성/실행/시각화



[그림 9] Notebook 안에 개별 Plugin을 사용하는 여러 Paragraph를 엮어 분석할 수 있다.

(출처: http://zeppelin.apache.org)

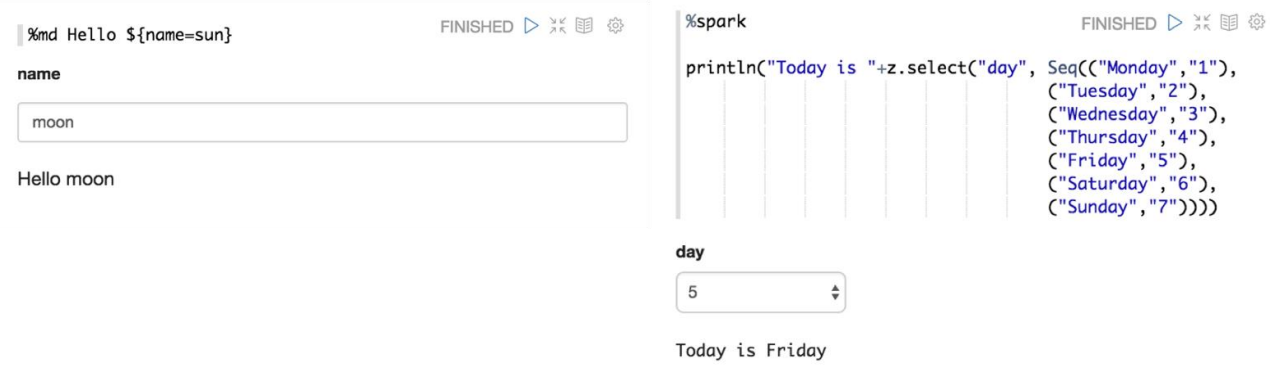
R이 그러한 것처럼 Hadoop과 Spark 역시 데이터를 불러들이고, 가공 및 변환하고 분석 알고리즘을 돌리는 과정을 코드로 작성해야 한다. 그리고 이의 결과를 여러 가지의 Graph를 통해서 시각적으로 보여주는 일 또한 해야 하는데 Zeppelin에선 이러한 작업을 단일 Notebook 안에서 할 수 있다. 그것도 자신이 선호하는 Framework 또는 언어를 선택할 수 있다. 방법은 Paragraph를 하나 생성한 후 %pyspark 처럼 어떤 플러그인을 사용하겠다는 선언을 해주고 그 아래 관련 코드를 작성하면 된다. 예를 들어, 인터넷에서 데이터를 받아 HDFS에 적재하고 이를 PySpark에 불러들여서 실행하는 작업을 해야 한다면 %bash 플러그인에서 wget 등을 데이터를 다운받은 후 %hdfs 플러그인으로 데이터를 HDFS에 적재하고 %pyspark로 데이터 처리를 하면 된다. 그리고 그 결과를 %sql로 정리한 후에 Paragraph의 우측 상단의 버튼을 클릭하여 결과를 막대/선/파이 차트 등으로 확인할 수 있다. 주의할 점은 Paragraph는 위에서 아래로, 좌에서 우로 실행되므로 이 부분만 주의한다면 다양한 기술 스택들을 서로 엮어 분석코드를 만들 수 있다. Notebook 전체를 한꺼번에 실행할 수도 있고, Paragraph 별로 개별적으로 실행할 수도 있으며 Notebook을 Cron에 물려 자동으로 실행하도록 할 수도 있다.

3.2 협업

IPython등과 차별화되는 Zeppelin만의 강력한 기능 중 하나가 바로 협업 기능이다. Google Apps를 보면 여러 사람이 같이 동시에 Google Sheet나 Google Document 등을 편집할 수 있는데 Zeppelin도 WebSocket을 활용하여 같은 Notebook을 여러 사람이 동시에 편집할 수 있는 기능을 제공하고 있다.

한 사람이 분석 코드를 짜면 다른 사람이 그 결과를 원격에서 실시간으로 바로 확인할 수 있다. 이러한 기능은 원격으로 공동 작업을 통해서 실험과 논문을 작성하는 일이 많은 연구소에서 환영할만한 기능이다. 다만 Google Apps처럼 문서의 히스토리까지 알아서 관리해주는 기능은 2016년 현재 제공되지 않고 있다. (Git을 이용하여 비슷하게 지원하는 방법은 존재한다.)

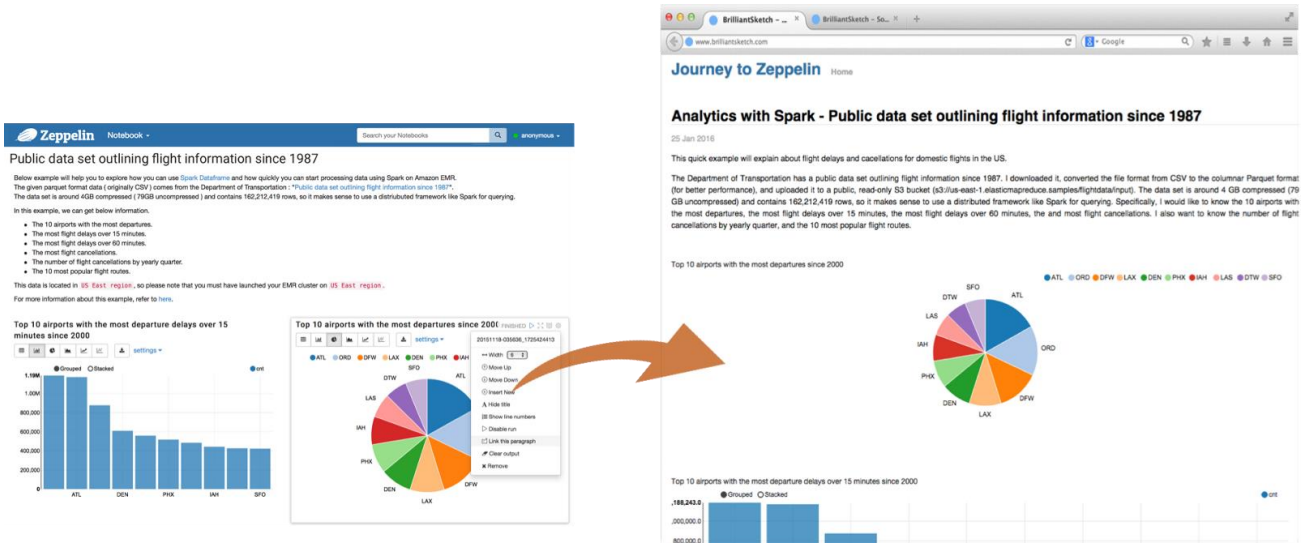
3.3 데이터바인딩



[그림 10] 분석 코드 안에 변수를 삽입하고 Binding을 걸어 Interaction이 가능하도록 만들 수 있다. (출처: http://zeppelin.apache.org)

이뿐만 아니라 Zeppelin은 Angular JS를 활용하여 Web Interface가 만들어졌기 때문에 Angular JS의 장점인 Data Binding 기능을 응용한 보고서 또는 Dashboard를 만들 수 있다. 예를 들어 학교에서 학생들의 성적을 분석하는 보고서를 만든다고 하자. 이 보고서에는 50점 이상 학생들의 분포만 보여주는 그래프가 포함되어 있고 이 그래프를 만든 이는 코드를 고쳐 50점을 60점으로 수정한 그래프를 쉽게 만들 수 있을 것이다. 하지만 보고서를 보는 사람이 코드를 잘 모르는데 40점이나 70점 등으로 이리저리 확인해서 보고 싶다면? Zeppelin에선 이 기준점수를 TextBox로 받을 수 있도록 Binding을 걸 수 있다. 이렇게 하면 코드를 모르는 사람도 보고서에서 점수만 고치면 그래프가 자동으로 변경된 값을 반영하여 새로 그려지므로 좀 더 사용자 친화적으로 보고서가 만들어진다. Binding뿐만 아니라, Angular 인터프리터를 사용하여 AngularJS가 가미된 HTML 코드를 실행하면 Zeppelin 소스를 수정하지 않아도 여러 GUI를 만들어 낼 수 있어 다양한 Dashboard를 구성할 수 있다.

3.4 공유 및 확장



[그림 11] 분석한 결과를 다른 웹사이트에 embed할 수 있다. (출처: http://zeppelin.apache.org)

IPython등은 데이터 분석/연구 결과를 정리해서 보여주기 위한 연구 노트 같은 기능제공에 초점이 맞춰져 있지만, Zeppelin은 초기부터 분석 결과를 Dashboard 형태로 여러 사람이 공유할 수 있게 하도록 하는 데 초점이 맞춰져 있다. Paragraph 우측 상단에 Export/Share 버튼을 사용하면 그 Paragraph를 IFrame을 이용하여 다른 Website에 Embedded 시켜줄 수 있다. 또한, %html, %md (markdown) 등의 자바스크립트 삽입 및 문서 스타일링도 지원하기 때문에 자신의 입맛에 맞게 버튼을 달거나 분석 결과를 다른 모양으로 표현하는 것도 가능하다. 이런 기본 기능들이 맘에 안 든다면 새로운 플러그인을 작성하거나 D3.js를 건드려서 나에게 맞는 형태로 수정해줄 수도 있다. 물론 다소 빌드가 까다롭기 때문에 이 부분은 현재는 쉽진 않지만, 현재 시험적용 중인 Helium Application 기능이 안정화 되면 좀 더 확장성이 좋아질 것이라 기대한다.

4 맺음말

[그림 12] https://goo.gl/PDJuyM 또는 페이스북 "제플린과 친구들"로 가입하실 수 있습니다.

아마존, 마이크로소프트, 트위터, 페이스북 같은 글로벌 IT 공룡뿐만 아니라 국내에선 VCNC(비트윈), 캐시슬라이드 같은 스타트업부터 SK 같은 대기업까지 활용하고 있는 Zeppelin은 데이터 분석 플랫폼을 개발하는 NFLABS(현재 ZeppelinX)라는 한국 스타트업에서 만들었고 한국인들이 Maintainer로 활동하며 Apache Software Foundation의 지원을 받는 전 세계에서 몇 안 되는 한국인이 주도하는 오픈소스 프로젝트이다. 한국인 PMC 분들을 중심으로 국내 Zeppelin 유저들을 지원하기 위해서 페이스북 그룹을 운영하고 있는데 이곳에서 비정기적으로 다양한 교육이나 번개 모임 등을 하고 있으니 사용하시는 분이나 관심 있으신 분은 이곳에서 Zeppelin을 만들고 계신 분들과 바로 소통하셨으면 좋겠

다.