< 요 약 서 >

1. 과 제 명	공개SW 개발지원사업(지정과제: 빅데이터 분석 및 추천 기술 개발)		
2. 총사업기간	2015. 5. 1. - 2015. 11. 30. (7개월)	**3. 총 투입인원**	총 7 명
4. 사업비	총연구비: 150,000천원		
	정부출연금: 150,000천원, 민간부담금: 0천원		
5. 참여기관			
6. 공개SW 라이선스	Apache License version 2		
7. 개발 목표	Apache Zeppelin (Incubating)에서 데이터 처리 프레임워크 연동모듈 개발		

Apache Zeppelin (Incubating) 프로젝트는 Hadoop 을 중심으로 하는 big data 생태계에서 데이터 시각화를 포함한 분석환경을 제공해주는 툴입니다. 오픈소스 프로젝트로 NFLabs에서 시작이 되었고 2014년 12월 Apache 소프트웨어 재단의 인큐베이팅 프로젝트로 채택되었습니다.
Zeppelin 은 본 과제를 통해 데이터 처리 프레임워크 연동 모듈을 추가로 개발 하는것을 목표로 합니다. 새로운 데이터 처리 프레임워크 연동 모듈개발을 통해 기존 연동 모듈과 함께 Hadoop 생태계의 거의 모든 데이터 처리 프레임워크들을 사용할 수 있게 되어 Zeppelin 이 표준 빅데이터 분석 환경으로 자리매김 할 수 있게 합니다. 연동 모듈 개발을 통해 새로운 기능을 포함한 릴리즈를 수행함으로서 인큐베이팅 프로젝트에서 Top Level 프로젝트로 승격되는데 도움이 되며, Top Level 프로젝트가 되면 더 욱 많은 되면 사용자와 개발자 커뮤니티에서 프로젝트의 영향력과 신뢰도가 크게 향상되고, 다른 아파치 재단의 소프트웨어와의 공생관계를 구축하여 big data 시장에서 분석 환경으로서 한축을 담당하게 될 것입니다.

8. 개발내용

Zeppelin 은 다양한 데이터 처리 프레임워크를 플러그 인하여 분석을 수행할 수 있도록 설계되어있습니다. 현재 Zeppelin 은 일반 분석 엔진인 Apache Spark 과 SQL on Hadoop 시스템인 Apache Hive, Apache Tajo 와 연동모듈이 개발되어 있습니다.

기존 개발된 연동 모듈과 더불어 다음의 데이터 처리 프레임워크와의 연동 모듈을 개발합니다.

* Apache Flink (http://flink.apache.org/)
* Apache Lens (Incubating, http://lens.incubator.apache.org/)
* Apache Ignite (Incubating, http://ignite.incubator.apache.org/)
* Apache Geode (Incubating, http://geode.incubator.apache.org/)

각각의 연동모듈은 데이터 처리 프레임워크에 따라 내부적으로 하나 또는 그 이상의 연동모듈의 집합으로 이루어집니다.

9. 과제수행방법

연동 모듈을 개발하기 위해서는 각각의 데이터 처리 프레임워크 프로젝트와의 협업이 필요합니다. 따라서 다음과 같은 절차를 통해 연동 모듈을 개발하게 됩니다.

1. 각각의 데이터 처리 프레임워크 커뮤니티에서 Zeppelin과 연동할 수 있는 방법에 대한 토론
2. 토론 결과를 바탕으로 Zeppelin 프로젝트에 이슈 생성
3. 주관사의 개발인력과 커뮤니티를 통해 참여하는 오픈소스 컨트리뷰터가 같이 개발을 수행
4. 개발한 코드를 테스트하고 새로운 버전을 릴리즈
5. 개발에 활발히 참여한 컨트리뷰터를 커미터로 선출

10. 결과활용 및 사업화 계획

Apache Zeppelin (Incubating) 은 이미 Apache Spark 을 사용하는 사용자/기업에서 사용하고 있습니다. Apache Spark과 함께 big data 분석 pipeline 을 구성하는데 아주 유용하게 사용할 수 있습니다. 특히 데이터 분석가에게는 data의 ETL, 탐색, 분석, 시각화 까지의 과정을 하나의 환경에서 제공함으로 생산성 향상에 큰 도움이 됩니다.

현재 국내 game 사, portal, sns 서비스 제공사, mobile app 서비스회사 등에서 데이터 분석을 위해 유용하게 활용되고 있습니다. 또한 Big data software stack 을 패키지하여 사업화하는 글로벌 회사중에 하나인 hortonworks에서 자사의 제품인 HDP에 Zeppelin 을 탑재하려 하고 있습니다.

NFLabs 에서는 Apache Zeppelin (Incubating) 의 사용자 저변을 토대로, Apache Zeppelin (Incubating)의 기술지원, Apache Zeppelin (Incubating) 을 더 유용하게 사용할 수 있는 부가 서비스 개발을 통해 사업화 할 계획입니다.

11. 최종결과물

최종 결과물은 Apache Flink, Apache Lens, Apache Ignite, Apache geode 와의 연동모듈 (소스코드) 가 결과물로 발생됩니다.

결과물을 내기위해서 다음과 같은 과정을 거치게 됩니다.
- 새로운 committer 충원
- 2번 이상의 버전 릴리즈
- Apache Top Level 프로젝트로의 승격

12. 경제적파급효과

현재 Big data 산업을 주도하고 있는것은 Apache 소프트웨어 재단의 Hadoop 생태계입니다. Hadoop 생태계는 여러 글로벌 기업들의 자사 big data 솔루션의 근간이 되고 있습니다. IBM, Oracle, Cloudera, Hortonworks, Pivotal, MapR 등 거의 모든 big data 솔루션 제공업체들은 Hadoop 생태계에 의존하고 있습니다. 이러한 Hadoop 생태계에 다양한 소프트웨어 스택이 있습니다. 스토리지(HDFS), 컴퓨팅(Map-Reduce, Spark, etc), SQL-on-Hadoop (Hive, Tajo, Drill, MRQL..), NoSQL (Hbase, accumulo, Cassandra, etc)
하지만 분석 환경을 제공하는 소프트웨어 스택이 현재로서는 Apache Zeppelin (Incubating) 이외에 존재하지 않습니다.

Big data 시장이 Hadoop 생태계에 의존하고 있고, Hadoop 생태계는 아직 Zeppelin 이외에 분석환경에 대한 대안이 없는 점으로 Apache Zeppelin (Incubating) 이 Hadoop 생태계, 더 나아가서는 big data 시장의 사실상의 표준 분석 환경이 되고자 합니다.

일반적인 분석환경의 장악하고 있는 Excel 이 수십년간 Microsoft 사의 Operating System을 사용하게 만드는 killer app 이었을 정도로 파급력이 있었습니다. Excel 은 빅데이터를 처리할 수 없는 바, Zeppelin 이 big data 시장의 사실상 표준 분석 환경이 된다면, 향후 수십년간 Excel 이상의 파급력을 가지게 될것으로 생각됩니다.