Flink Notebook

*Interactive power of Flink Streaming*



Spring 2015

COMMENTS ON 12th APRIL

Please, create a standard EIT ICT LAbs cover page,   
see at TEMPLATES folder

DO NOT PUT LEAVE A SPACE BEFORE A COMMA.

Please, correct page numbering

Flink Notebook

*Interactive power of Flink*

1. **Introduction (2-3 pages)**
2. **The business idea**

Hint: online processing of streaming data needs a new solution, a compiler ; the Thesis is offering a PROJECT to develop this tool  
 *terms, short idea*

// Big data area

In the last few years Big Data HAVE A SENTENCE EXPLANATION OF “BIG DATA” IN FOOTNOTE!! THE THESIS IS TARGETED FOR BUSINESSMEN, FOR MANAGERS! generated a lot of buzz along with the launch of several successful big data products. Thanks to contribution from open source community and several giant Internet companies, the big data ecosystem has now reached a tipping point, where the basic infrastructure capabilities of supporting big data challenges and opportunities are easily available. Entering the next generation of big data - so-called Big Data 2.0 - two of its concentrated areas are Velocity and Applications, besides Data Quality. The cause for the former is that data is growing at an exponential rate and the ability to analyse it faster is more important than ever. For instance, sensors can generate data on millions of events per second and store all of those data and response in real-time is non trivial. The latter is helping to overcome the technical challenges of existing frameworks by making them easy to use and understand for everyone to benefit from big data.

// demand for a streaming solution

As a result, the demand for ***streaming processing*** USE HIGHLIGHTING FOR IMPORTANT TERMS, WORDS! is increasing a lot these days. Processing big volume of data is not sufficient in the cases that infinite streaming data is arriving at high speed and ~~required~~ users require a system to process fast and react to any incident immediately. In addition, although hardware price has plunged year over year, it’s still expensive to equip a storage which is growing few terabytes every day for batch analysis. Streaming processing solutions are designed to handle high volume in real time with a scalable , high-availability and fault tolerant architecture. This enables analysis of data in motion.

// the disadvantage of Streaming framework (Customer pain)

One of the disadvantages to users is that many big data framework provide a rich imperative API codes only to process data stream on the fly. First, user must spend time to learn API documentation properly since those APIs is fairly new to them. Therefore, the life cycle to develop products taking longer. Second, given that most big data applications are fairly simple application-wise, a block of API codes might be less optimal to use for most the popular queries. Third, the only way to share your work is to pack it as a library or service which is need to be deployed again. Fourth, the results of streaming queries are keep changing upon the time. Thus, it demands a visual way to monitor the streaming instead of concrete array of numbers. Those above drawbacks apparently provide an unfriendly UX that holdback both productivity and system performance. It is really against what we expect from “Application” aspect of Big Data 2.0.

INSERT HERE MAYBE A SIMPLE IMAGE, LIKE ATTACHED OR FROM  
<http://www.intelligentbusiness.biz/latest-news/real-time-analytics-route-continuous-risk-management-business-optimisation/>

ETC. TO DEMONSTRATE THE PROBLEM!

// Statement about Flink Notebook.

We would like to promote “Flink Notebook” , a interactive, collaborative web-based interface on top of Apache Flink which provide a complete streaming processing. Flink Notebook brings TO A developer a friendly, effective environment without scarifying the power of Flink Streaming.

Basically, Flink Notebook allows users to write standard SQL queries on webpages. The notebook issues corresponding commands to Flink to make it operate continuously on data as they arrive. The result is updated incrementally in real-time and displayed on Notebook with visual aid. In short, users need to know nothing about Flink API, except for very common SQL basis. The result is also brought to life by a powerful visualisation web pages.

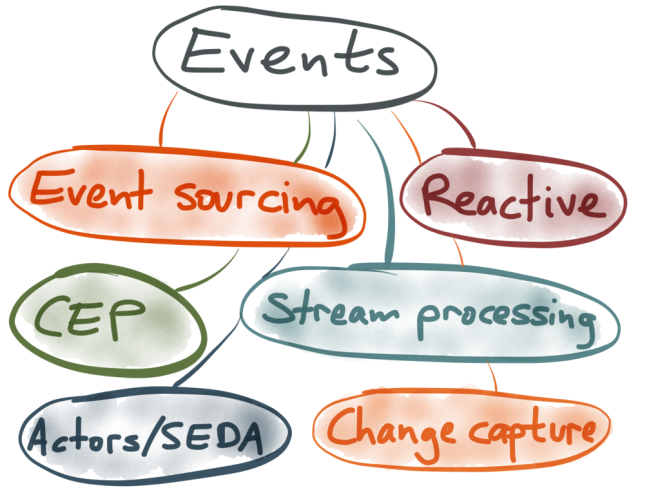
Flink Notebook facilitates the productivity by allowing several colleagues to collaborate on the same notebook on the same work in real-time despite the physically location differences. The work can also be captured and exported to a portable format which can be viewed shared , restored later at ease.

GOOD TEXT, EASY TO UNDERSTAND, CLEAR

1. **My research background**

Hint: just very shortly, why is it important, what happened before

“During my internship I am working on …. Part of this topic is related to online, streaming data processing and using this idea my Minor Thesis concept is to propose a company project building a new software tool.“ MAYBE YOU CAN START THIS SECTION WITH A SENTENCE LIKE THIS!

Streaming Processing is not a new concept. Indeed the similar concept, Complex Event Processing ~~(CPE)~~ CEP had been proposed from the 1990s by Event Simulation Research at Stanford [1]. Since that time, people have started generating a lot of different buzzwords around it [2] and often reinventing ideas borrowing from other fields , but using a different vocabularies to describe the same concepts. Basically , the idea is to analyse one or multiple data streaming to identify meaningful events and respond to them as quickly as possible. 

OKAY, I SEE A GOOD PICTURE WHICH MAKES UNDERSTANDING EASIER.

According to CEP Tooling Market Survey 2014 [3], from 1996, there were more than 30 companies providing Streaming Processing solution. All the major software vendors (IBM, Oracle, Microsoft, SAP) also have good to excellent offerings in the CEP space for customers.

However, since a massive amount of data is growing rapidly every second, Hadoop is emerging distributed processing ecosystem today. INSERT A FOOTNOTE HERE TO EXPLAIN Hadoop LIKE “Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It is part of the [Apache](http://searchcio-midmarket.techtarget.com/definition/Apache) project sponsored by the Apache Software Foundation.”

Thanks to Hadoop, people can build a large scalable distributed system on Cloud. Even though Hadoop is designed to scale up to thousands of machines with very high degree of fault tolerance, it is optimised to run batch jobs with a huge load of computation. Because of time factor, Hadoop has limited value in online environment where fast processing is crucial. Therefore, existing CEP solution is barely compatible with Hadoop ecosystem. We demand a new sort of streaming framework which is able to integrate on top of Hadoop system. Apache Flink is one of those frameworks ????? THESE FRAMEWORKS.

In term of cooperative work, iPython Notebook is web-based interactive environment where one can combine a working session into a single document. The document can be shared AND to BE converted to other common formats such HTML, pdf for quick review. Google Docs allows multiple users to collaborate on a document in real-time. Our Flink Notebook inherits both of ~~those~~ these technologies to provide an interactive, collaborative environment along with super power of Flink

REFERENCES CAN BE ONLY AT THE END OF TEXT. TRY TO USE AUTHOR+TITLE (IF EXISTS: DATE) BESIDE THE PURE URL !

[1]<http://complexevents.com/stanford/rapide/>

[2] http://blog.confluent.io/2015/01/29/making-sense-of-stream-processing/

[3] http://www.complexevents.com/2014/12/03/cep-tooling-market-survey-2014

1. **Method to use in the Thesis**

Hint: very shortly: literature review, some interviews, company consultations,   
my own major thesis research to use

// literature review

// interview with Developers (SQL syntax design)

// open source project

YOU SPEAK HERE ABOUT A PROPOSED PROJECT. USE PROPER WORDS TO MAKE IT CLEAR THIS IS ONLY A PORJECT PORPOSAL: HERE YOU WILL NOT DEVELOP ANYTHING!

A project developing Flink Notebook ~~required~~ REQUIRES us to have a deep understanding on Flink Streaming , interpreter theory and web programming skill. First, we study the principle of Stream Processing and data flow in Flink. Second, we need to design SQL-like syntax which user can use to issue command to Flink. Third, a web application helps to interact with user.

The project involves many big data developers during design process. According to regular project management methods, we have to ~~We~~ observe, to monitor their behaviour during testing , also we need to interview them for valuable feedback. Finally, to shorten the developing cycle, we take the advantage of community to integrate many open source project into ours.

1. **Marketability of the Idea**
2. **The product environment (2-4 pages)**

Hint:

literature research on compilers / streaming data and referring to the major thesis;

Software development methodologies, comparison to these;   
1-2 GENERAL articles and 2-3 SPECIFIC ones to this topic;  
have an opinion summarised about these articles

See some articles, reports attached: just place a short history plus definitions here; referring even to common textbooks

* Data Stream: what is data stream
* CQL : what is CQL
* Compiler in general

Flink , SQL syntax

1. **Market needs**

Hint: existing customer problems statistical data, how many project, programmers, etc could be interested „I just focused in the Thesis to the Vietnamese market: product specialities   
can be used at approximately XXX. Software development companies in VietNam

SIMPLE SENTENCES:   
“ACCORDING TO WELL-KNOWN Forbes Magazine forecast …  
<http://www.forbes.com/sites/gilpress/2014/12/11/6-predictions-for-the-125-billion-big-data-analytics-market-in-2015/>

refer e.g. to simple market analysis like  
<http://onlinelibrary.wiley.com/doi/10.1002/bltj.20528/abstract;jsessionid=2BF59F4BBD17C34AD663A78A74540F6E.f02t03?deniedAccessCustomisedMessage=&userIsAuthenticated=false>

<http://www.reportlinker.com/report/best/keywords/Software?utm_source=adwords1&utm_medium=cpc&utm_campaign=Business_Services&utm_adgroup=Software&gclid=CP-L1tXb8cQCFQzJtAodnjoA0Q>

<http://marketinfogroup.com/north-americas-saas-market-to-top-85-billion-in-the-next-five-years/>

The figures are broken down by the following segments:

Software (incl. SaaS):

Data Infrastructure

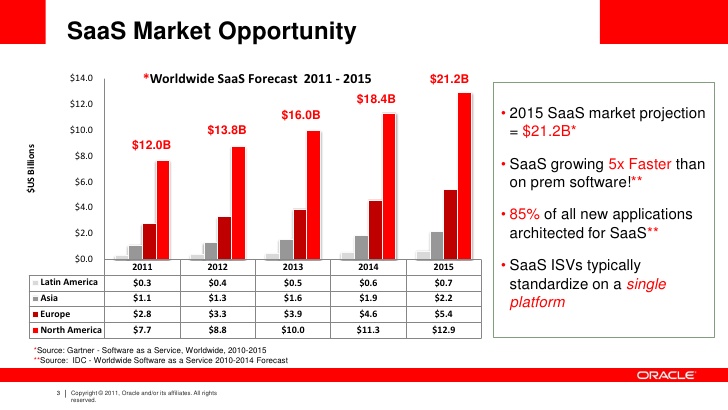
Applications & Analytics

IT Services:

Data Infrastructure (Consulting, SI and AM)

Applications & Analytics (Consulting, SI and AM)

Hosting & Support Services



SOURCE: http://www.slideshare.net/cvidya11/building-a-scalable-profitable-saas-business-13282375

// focus on market in Germany and Hungary

// number of programmers, big data company here

DO NOT BUILD A VERY COMPLICATED SITUATION FOR YOURSELF: IF THERE IS NO DATA ON SOMETHING, LEAVE IT OUT!

1. **Competitors**

Hint: List and put into a matrix all known parameters of leading products

1. **Value proposition**

Hint: what we can offer   
We would offer a difference to the above products; it will be a   
- compiler, easy-to-use; more effective, etc etc  
- a „Notebook” to use as a manual

1. **The project proposal**
2. **The project idea**

Hint: Repeating in details the starting text: why is it beneficial for the company to have this development

1. **Customer segment, targeted markets**

***Hint:*** *talk about „two specific markets, ~~Italy,~~ GERMANY VietNam”  
or number of small developers, etc etc as you like*

1. **~~[deleted] Channels to build a user community on these markets~~**
2. **~~[deleted] Contacting customers: customer relationship solution, promotion~~**
3. **The company background**

Hint: here you can write about the company, today product portfolio, etc  
why they are interested in a new product

// DataArtisan

Very few sentences, maybe a corporate organizational chart to understand where to “place”

your project, what do you propose : which type of people from the organization could take part in the project?

1. **Feasibility**

Hint: Activities and resources needed for the project  
SOME SENTENCES ABOUT PROJECT MANAGEMENT; ICT projects;   
development process: resources (mandays, experts, software to use,   
hardware needs, time, money, etc.  
CREATE A TABLE, A MATRIX HERE FOR SAY, 4 QUERTERS (ONE YEAR PROJECT), WITH ACTIVITIES, KEY FIGURES  
Marketing / promotion activities: building a website; pricing the tool  
compared to other tools, etc

1. **Project planning**

Hint: A simple scheduling of activities (based on list of 3.6.); participants, milestones and deadlines;  
all in a visual plan like Gantt-diagram  
Management problems: which type of project management is needed

1. **My participation in the proposed “Customer Feedback Subproject Proposal”**

Hint:

A simple scheduling of activities (based on list of 3.6.); participants, milestones and deadlines;  
all in a visual plan like Gantt-diagram  
Management problems: which type of management is needed

1. **~~[Delete] Building a brand~~**

~~Hint: however being small, any possibility to use a partner to be a „branded product”~~

1. **~~[Revenue Model]~~**

~~Hint: pricing ideas: there will be a lecture about; building a brand~~

1. **Summary**

Hint: What was the target; how I worked, what is the result;  
recommendations to many internship company management;  
and: What I have learnt form this Thesis Work…

———— IGNORE IT -----------------------------

<http://www.bi.hu/index.html>

<http://www.bi.hu/2015/03/30/a_bi_platformok_vezeto_szallitoi_2015ben.html>

<http://bievkonyv.hu/2011/evkonyv.shtml>

<http://bievkonyv.hu/2011/konferencia.shtml>

<https://www.pac-online.com/big-data-segments-market-figures-hungary-0>

<http://www.fujitsu.com/hu/solutions/business-technology/bigdata/>

———————-

When I was working on Google Apps, we would often hear people ask, “Why launch Google Spreadsheets? It’s 20 years behind Microsoft Excel and 200 features short!” They didn’t realize that a driving mantra for Google Apps was “It’s the collaboration, People!” I have seen metrics, and still experience daily, how Google Apps’ real-time collaboration features boost team task productivity by a factor of 10x or more. It is collaboration among team members with diverse skillsets and points of view that yields these large gains in organizational smarts.

———————-

Company:

Companies

actuated

Alteryx

Applix

Business Objects

Cloudera

Cognos

Comshare

DatAllegro

EMC

Eclipse BIRT

Greenplum

HP

HortonWorks

Hyperion

IBM

Infobright

Informatica

JasperSoft

Jedox Palo

KarmaSphere

MicroStrategy

Microsoft

Netezza

, Oracle

Panorama Software

ParAccel

Pentaho

QlikView

Rapid Miner

SAP

SAS

SPSS

Splunk

Sun

Sybase

TIBCO Spotfire

Tableau

Talend

Teradata,

Vertica

**MODIFIED** Thesis frame DUY  **22nd March 2015**

1. Introduction

1.1. The business idea: online processing of streaming data needs a new solution,   
a compiler ; the Thesis is offering a **PROJECT** to develop this tool  
 *terms, short idea*

1.2. My research background  
*just very shortly, why is it important, what happened before*

1.3. Methods to use in the Thesis  
*very shortly: literature review, some interviews, company consultations,   
my own major thesis research to use*

2. Marketability of the idea

2.1. The product environment: literature research on compilers / streaming data   
*and referring to the major thesis;* Software development methodologies, comparison to these;  *1-2 GENERAL articles and 2-3 SPECIFIC ones to this topic;  
have an opinion summarized about these articles*

2.2. Market need: existing customer problems  
*statistical data, how many project, programmers, etc could be interested  
„I just focused in the Thesis to the Vietnamese market: product specialities   
can be used at approximately XXX. Software development companies in VietNam*

2.3. Competitors: similar products and services  
List and put into a matrix all known parameters of leading products

2.4. Value proposition: what we can offer   
We would offer a difference to the above products; it will be a   
- compiler, easy-to-use; more effective, etc etc  
- a „Notebook” to use as a manual

3. The ~~Business Model~~ Project Proposal

3.1. The Project Idea  
Repeating in details the starting text: why is it beneficial for the company to have this development

3.2. Customer segment, targeted markets  
*here you can talk about „two specific markets, Italy, VietNam”  
or number of small developers, etc etc as you like*

~~3.3. Channels to build a user community on these markets~~

~~3.4. Contacting customers: customer relationship solutions, promotion~~

3.5. The company background  
*here you can write about the company, today product portfolio, etc  
why they are interested in a new product*

3.6. Feasibility: Activities and resources needed for the project  
*SOME SENTENCES ABOUT PROJECT MANAGEMENT; ICT projects;   
  
development process: resources (mandays, experts, software to use,   
hardware needs, time, money, etc.  
Marketing / promotion activities: building a website; pricing the tool  
compared to other tools, etc*

3.7. Project Planning   
A simple scheduling of activities (based on list of 3.6.); participants, milestones and deadlines;  
all in a visual plan like Gantt-diagram  
Management problems: which type of management is needed

3.8. My participation in the proposed „Customer Feedback Subproject Proposal”   
*Here you can write some words on what you could do: idea planning, programming, testing, presentation to beta-version users, etc*

~~3.9. Building a brand MAYBE~~*~~however being small, any possibility to use a partner to be a „branded product”~~*

~~3.10. Revenue model~~*~~pricing ideas: there will be a lecture about; building a brand~~*

~~3.11.~~

4. Summary  
*What was the target; how I worked, what is the result;  
recommendations to mny internship company management;  
and: What I have learnt form this Thesis Work…*