**Project Requirement**

This project is aimed at practicing the big data applications and analytics. A recommendation service system should be established to allow users navigating the large amount of data, and to make decisions based on the recommendations. This system manages all of its data at the back-end and supports browsing and searching interfaces over the Internet. These functionalities could be provided by integrating tools.

Each team will select the topic and distribute the workload based on the team decision.

All parts of the project should be done either individually or by a team of students. Teams will be established prior to the beginning of team assignments. Each team may have 4-8 people. Since the teamwork is the cornerstone of practical IT service world, it is important that you experience being a member of a virtual team in this collaborative learning environment.

Programming is NOT the focus of this project which should be linked directly to the user perspective/experience, or service-oriented tasks. In other words, as long as the system is available, useful, and effective, it does not matter on using whatever programming skills or languages.

**Due time:** Target to the mid-night of last class day.

该项目旨在实践大数据应用和分析。 应建立推荐服务体系，允许用户浏览大量数据，并根据推荐做出决策。该系统在后端管理其所有数据，并支持通过互联网浏览和搜索界面。这些功能可以通过集成工具提供。

每个团队将根据团队决策选择主题并分配工作负载。

项目的所有部分都应单独或由学生团队完成。团队将在团队分配开始之前建立。 每个团队可能有4-8人。 由于团队合作是实际 IT 服务世界的基石，因此在这个协作学习环境中体验成为虚拟团队的一员非常重要。

编程不是这个项目的重点，它应该直接链接到用户视角/体验，或面向服务的任务。 换句话说，只要系统可用、有用且有效，使用任何编程技能或语言都无关紧要。

到期时间：目标到上课日的午夜。

**Project Parts**

Tentatively, this project could be broken down into five parts:

**Part 1**: Project Plan (15%)

A project plan should be generated to contain the necessary information as an IT project, such as project target, motivations, resource estimation, schedule arrangement, and etc.

**Part 2**: Data transformation (15%)

A volume of data should be collected and formatted. The data should be examined and designed as a data set for further utilization of big data analytics.

**Part 3**: Recommendation system (30%)

A recommendation service system should be designed and set up. The service system is the main development and testing platform. Once the service system is setup and running, users will be able to interact with back-end to obtain recommendations based on big data analytics.

**Part 4**: Implementing service interface (15%)

A service interface should be available to allow users interact with the recommendation system. The service should support queries both on numeric attributes (like birth day) and textual attributes (like skill/personality description) and allow simple keyword-based searches on textual attributes.

**Part 5**: Documents (25%)

Finally, the documents should be generated to wrap up the project and to guide users through a training procedure. The following items may be helpful:

1. The system architecture diagrams and configurations.
2. The user guides or operational menus for the service system.
3. The potential training process/materials, such as demo, screenshots, or PowerPoint presentation.
4. The source code and/or executable package.
5. The report of project team work, including 1) name list team member; 2) task distribution for each individual, 3) ***amount of work for each role contribution***.

暂定，该项目可分为五个部分：

第 1 部分：项目计划 （15%）

应生成项目计划以包含作为IT项目的必要信息，例如项目目标，动机，资源估算，进度安排等。

第 2 部分：数据转换 （15%）

应收集大量数据并格式化。 应将数据检查并设计为数据集，以便进一步利用大数据分析。

第3部分：推荐系统（30%）

设计建立推荐服务体系。 服务体系是主要的开发和测试平台。一旦服务系统建立并运行，用户将能够与后端交互，以获得基于大数据分析的建议。

第 4 部分：实现服务接口 （15%）

服务接口应可用，以允许用户与推荐系统进行交互。该服务应支持对数字属性（如出生日）和文本属性（如技能/个性描述）的查询，并允许对文本属性进行基于关键字的简单搜索。

第 5 部分：文档 （25%）

最后，应生成文档以结束项目并指导用户完成培训程序。 以下项目可能会有所帮助：

系统体系结构图和配置。

服务系统的用户指南或操作菜单。

潜在的培训过程/材料，例如演示、屏幕截图或 PowerPoint 演示文稿。

源代码和/或可执行包。

项目组工作报告，包括1）名单组成员名单;2）每个人的任务分配，3）每个角色贡献的工作量。

**Notes to Students**

**Selection of Partners**

Students may implement the project individually or in teams of 4 – 8 people. The choice is up to each student, but please keeps the following rules in mind when you select your project partners:

An identical amount of work is expected and the same grading scale is used for individual and team projects. Faculty experience indicates that in general it is not necessarily easier or more productive to work in teams - it's largely a matter of personal preference and working style.

If you work in a team, choose your partners carefully. Teams are permitted to "divorce" at any time during the course (due to incompatibility, one partner dropping the course, or any other reason), and individual students may choose to team up as the project progresses, however students from divorced teams may not form new teams or join other teams. Put another way, if a student turns in any part of the project as part of a team, every later part of the project must be turned in individually or as part of the same team.

We will not entertain any complaints of the form "I did all the work and my partner did nothing." Choose your partner carefully!

If you work in a team, your work must be turned in jointly, as ONE submission. If we receive submissions from students in a team separately, they would get 10% off as a penalty. In addition, if the above separated submissions are different, the submission with a lower score prevails. Teamwork turned in as individual work will be considered as plagiarism and handled through official University channels.

**Academic Honesty**

Each of you is expected to submit your own original work, or the original work of your team in the project. On many occasions it is useful to ask others (the instructor, the TA, or other students) for hints or debugging help, or to talk generally about programming strategies. Such activity is both acceptable and encouraged, but you must indicate any assistance (human, online, or otherwise) that you received. Any assistance received that is not given proper citation will be considered plagiarism. In any event, you are responsible for coding, understanding, and being able to explain on your own or as a team all project work that you submit.

We will pursue aggressively all suspected cases of plagiarism, and they will be handled through official University channels. If you have any questions about this policy or about the degree to which we will pursue academic honesty violations, please discuss your concerns with the course instructor immediately.

选择合作伙伴

学生可以单独实施该项目，也可以以4-8人的团队实施。选择取决于每个学生，但在选择项目合作伙伴时请记住以下规则：

预期工作量相同，个人和团队项目使用相同的评分量表。教师的经验表明，一般来说，在团队中工作不一定更容易或更有效率 - 这主要是个人喜好和工作方式的问题。

如果你在一个团队中工作，请仔细选择你的合作伙伴。团队可以在课程期间的任何时间“离婚”（由于不兼容，一个合作伙伴放弃课程或任何其他原因），个别学生可以选择随着项目的进展组队，但离婚团队的学生不得组建新团队或加入其他团队。换句话说，如果学生作为团队的一部分上交项目的任何部分，则项目的后期每个部分都必须单独或作为同一团队的一部分上交。

我们不会接受任何“我做了所有工作，我的伴侣什么也没做”形式的投诉。仔细选择您的合作伙伴！

如果您在团队中工作，您的工作必须作为一份提交材料共同提交。如果我们分别收到团队中学生的提交，他们将获得 10% 的折扣作为罚款。此外，如果上述分开的提交不同，则以分数较低的提交为准。作为个人作品提交的团队合作将被视为抄袭，并通过大学官方渠道处理。

学术诚信

你们每个人都应该提交自己的原创作品，或者你的团队在项目中的原创作品。在很多情况下，向其他人（教师、助教或其他学生）寻求提示或调试帮助，或者笼统地谈论编程策略是很有用的。此类活动是可以接受和鼓励的，但您必须指出您获得的任何帮助（人工、在线或其他方式）。任何没有得到适当引用的帮助都将被视为抄袭。在任何情况下，您都有责任编码、理解并能够自行或作为一个团队解释您提交的所有项目工作。

我们将积极追查所有涉嫌抄袭的案件，并将通过大学官方渠道进行处理。 如果您对本政策或我们将追究违反学术诚信的程度有任何疑问，请立即与课程讲师讨论您的疑虑。

**Project Topic Samples** (used by US college students)

***Project Option 1***

***Stock Recommendation System***

***Motivation***

* To help stock buyers to make **wiser** choices
* To find those who are very good at gaining profits from stocks (experts)
* Using user-based collaborative recommendation to find the similarity between the buyer and the expert
* Recommend the stock buyer some stocks from the most similar expert.
* To ease stock buyers from the heavy burden of looking through thousands of stocks and making a wise choice.

***Dataset***

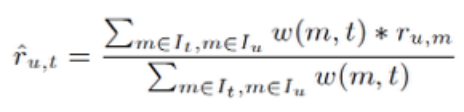
***NASDAQ Stock Exchange Data***

● Yahoo Finance dataset for historical prices

● Simulated trading records of 100000 users

**Algorithm**

● User-based Collaborative Filtering with Inferred Tag



**Tools**

* **Eclipse J2EE, Mahout, Maven, MySQL, PHP** server

**Progress**

○ Implemented UI and database

○ Determine the algorithm—User-based Collaborative Filtering with Inferred Tag Ratings.

**Expected Contributions**

○ Achieve improved Collaborative Filtering algorithm with Inferred Tag Ratings.

○ Present clients with sound stock recommendations based on the analysis of experts buying choices.

***Project Option 2***

*Movie Exploration*

***Motivation***

*Audience preference:*

I like comedies; I prefer blockbusters; I like Robert Downey Jr;

I’d like enjoy movies with my daughter; I ......

*Direction for Producers:*

What kind of movie customers would like to see?

Which actor is perfect match for this character?

Are there any qualities that make a memorable movie?

*Dataset:* online search

Tools: flexible

Algorithms:

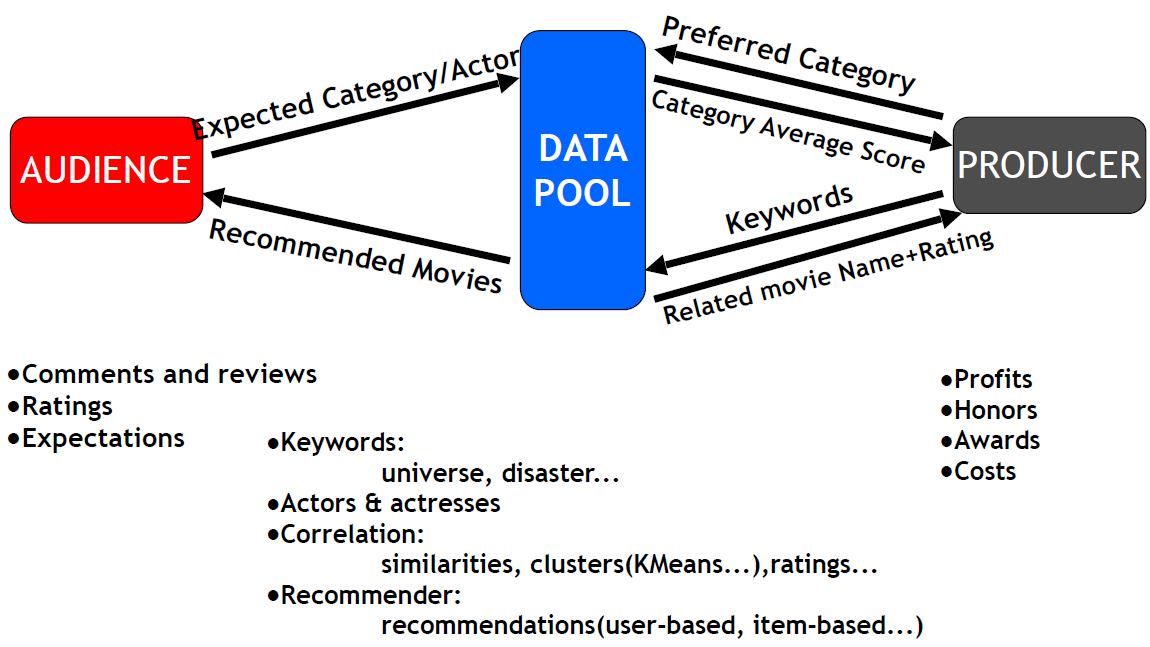
*Pig:* process user profile

Cluster & Classification:

Make correlation among audience

Recommendation:

Make predication based upon the keywords, extractions and preferences

***Expected Contributions***

***Project Option 3***

***Fantasy Basketball – Winning Strategy!***

***Motivation***

Sports play a big role among people in USA. As a result, people spend more time on watching NBA, MLB, NFL...etc. for their leisure time. Within the last five years, ESPN and Yahoo release an online sports game of all the professional sports league for people to actually have a chance to form a simulated team online and compete with other gamers. The ranking of the online teams is based on the real world behavior of the players they pick for their teams. Therefore, drafting becomes a key factor for the victory of the game. We wish to provide a winning strategy for drafting. Moreover, we hope to provide a practical dynamic strategy for real NBA team to choose players.

***Goal:***

* Recommend a player for the online gamer every round
* Cluster all the players into different groups of different functionalities such as scoring, rebounding…etc.
* Rank the players inside each clustered group
* Give suggestion based on the present status of the online gamers

***Dataset***

- NBA player’s statistics

<http://downloads.nbastuffer.com/nba-player-data-sets>

or on the bottom of

<https://www.nbastuffer.com/>

**Algorithms**

* Clustering
* Kmeans clustering, f-kmeans clustering…etc
* PLA algorithms, Linear Regression…etc
* Classification, Recommendation

**Tools**

* Mahout, Matlab, and every resource provided from class

***Progress***

- Survey on the point system of Fantasy Basketball

- Search for all the dataset that might be useful to our recommendation

- Get familiar with all the tools and algorithms that we might use

***Expected Contributions***

* Set up a dynamic recommendation system for the game
* Help the gamer to build a balanced champion team for the game

***Project Option 4***

*How to name your new-born baby (babies)?*

***Motivation***

In this project, something (relatively) simple, but very important in terms of answering a real-life question could be worked on.

What name(s) should you give to your new-born baby (babies)? I would look at historical popular male/female baby names, and use the various Recommendation, as well as Classification and/or Clustering techniques to help the expecting parents to make the right name choice, so their children do not complain about having not-so-cool names for the rest of their lives.

***Dataset***

Popular male/female baby names going back to the 19th Century from the Social Security Administration.

***Algorithms***

User-based and Item-based Recommendation, Clustering and/or Classification

***Tools***

R

HDFS/Hive/PIG/HBase

Mahout

***Expected Contributions***

* Set up a dynamic “name” recommendation system for the users

***Project Option 5***

***Exchange Rates Inquiry and Analysis***

***Motivation***

*With the international trade and commerce being of more* importance, it's necessary not only for a specific group of industries but also for all of us to gain daily access to the updated currency rates.

With more convenience in getting the exchange rate information, which hopefully will be provided by our project, people are equipped with more related knowledge to make better decisions regarding to the currency market.

Dataset:

❖ Instant exchange rates from Bloomberg

❖ History exchange rates from <http://www.livecharts.co.uk/historicaldata.php>

Algorithm:

• We would forecast the exchange rate of currency in some targeted countries against US Dollar in a developed market, applying scalable model to forecast in real-time. And we would like to use RMSE to measure the reliability and accuracy of our prediction. Besides, by showing the statistical significance, such as P values, the repeatability of the outcome will then be proved.

Tools:

❖ Eclipse

❖ Tomcat

❖ Apache

**Expected Contributions:**

• Our project is initially designed to provide users with the following contents.

• Forward and cross exchange rates for most world currencies

• Both instant and history data

• Basic analysis of exchange rates including regression and K means

• Latest news about exchange rates

**Key Factors Affecting Exchange Rate**

PPP(Purchasing Power Parity)

INT(Interest Rate Differential)—Such as Libor

GDP(The Difference in GDP Growth Rates)

IGP(Income Growth Rate)

Relative Economic Strength—we may use factors like GDP and IGP to measure it quantitatively

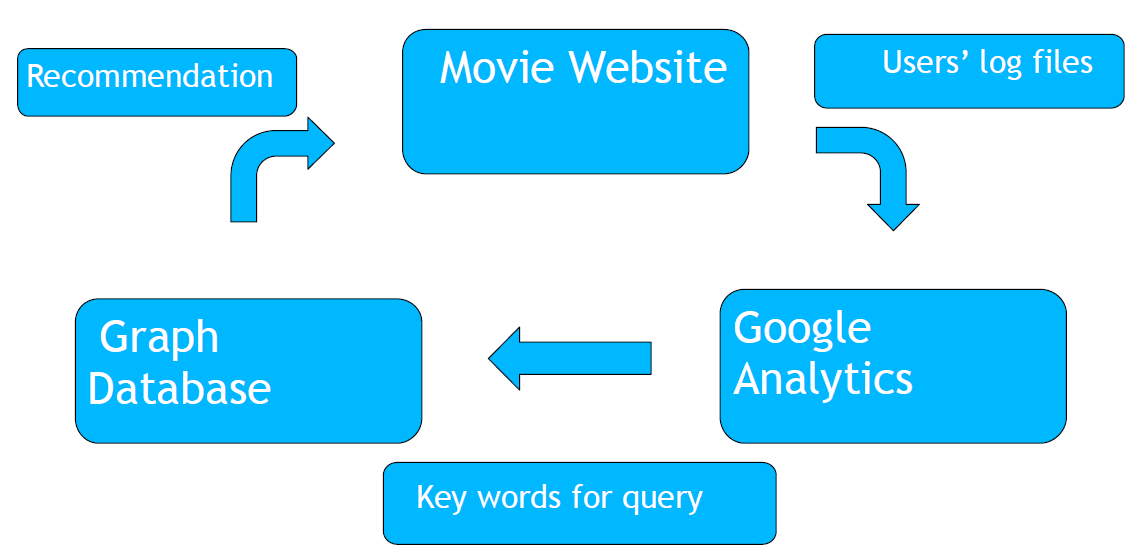
***Project Option 6***

*Google-Analytics, Graph – based Online Movie Recommendation System*

***Motivation***

- To build our own movie website and implement the movie recommendation functionality.

***Process Flow***

Dataset

MovieLens 1M dataset, which contains 1 million ratings from 6000 users on 4000 movies

- Algorithm

Various collaborative filtering algorithms, e.g. user-based recommendation, item-based recommendation etc.

- Tools

Web design - Dreamweaver / CoffeeCup

Users log file analysis - Google Analytics

Graph Database – Gremlin / Neo4j

Others – Mahout / Eclipse

Timeline:

Now -- 11/27/14: Movie website design and publish

11/28 – 12/04/14: Plug in Google Analytics in website and Analyze user log file

12/05 – 12/11/14: Generate Query for the Graph Database and update recommended movies on website

**Expected Contributions:**

* To build our own online movie website.
* To implement the movie recommendation functionality.
* To website analysis using Google Analytics

**Project Option 7**

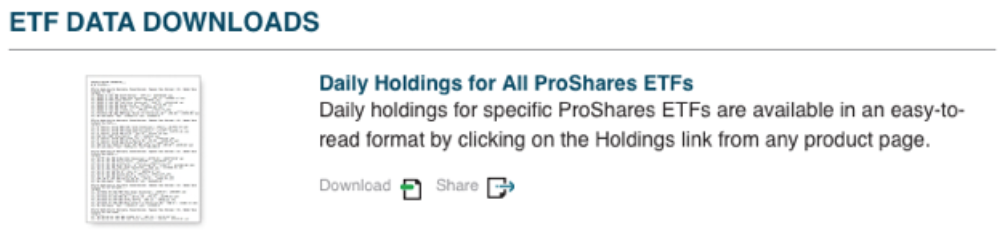
***STOCK FORECASTING USING HADOOP MAP-REDUCE***

**Motivation**

• Stock Market has high profit and high risk features, on the stock market analysis and prediction research has been paid attention by people. The stock price trend is complex nonlinear function so the price has certain predictability

• Hadoop MapReduce is a latest framework specially designed for processing large datasets on distributed sources. Apache’s Hadoop is an implementation of MapReduce

Dataset

Algorithms:

• Pearson Correlation Similarity

• Euclidean Distance Similarity

• Stochastic Gradient Descent (SDG)

Tools:

• Hadoop

• Mahout

• Hbase

**Expected Contributions:**

We are going to analyze the dataset called Daily Holdings for All ProShares ETFs which contains tons of information collected from the stock exchange market. The first step

is to scrutinize the data and provide the stock which may go up potentially. With these screened stocks, suggest a certain user a potential stock which she/he may be interested in.

***Project Option 8***

*Customer Complaint Analyses*

*Insights into Issues plaguing the Banking Sector*

***Motivation***

➢ One of the biggest challenges for banks is minimizing customer attrition rate which is directly dependent on customer satisfaction.

➢ Customers are inclined to choose the banks who can be trusted for their services.

➢ Banks make their decisions based on a subset of data because of absence of scalable solutions.

**In this project, a scalable design should be used to counter the above problems!**

***Dataset, Algorithm, Tools***

➢Consumer Complaints Database: The dataset contains

Retail consumer complaints with banks and financial institutions (provided by Consumer Financial Protection Bureau). http://www.consumerfinance.gov/complaintdatabase/

➢Algorithms: Various Clustering and Classification Algorithms

➢Tools and Languages: Hadoop, Mahout, Java, Python

***Expected Contributions***

➢Major retail banking issues by state and match-analyze them based on geographic or socio-economic brackets.

➢Top concerns of consumers in various states.

➢Derive business impact of customer satisfaction or dissatisfaction with their complaints on the institution.

➢Propose likely solutions that can be deemed as “first response” for future complaints of similar nature.

➢Hypothesize a performance metric to apply to all complaints – can be used to prioritize complaints based on resolution time.