BIG DATA ANALYSIS
– A CASE STUDY

Supervisor: Dr. Wang, Qing
Du, Haoran U5044502
OUTLINE

- Introduction
- Motivation
- Goal of the project
- Data structure
- Methodology/Techniques
- Experiment
- Challenges
- Q&A
INTRODUCTION

- Industrial project from the Piction company
  - The Piction Digital Media eXchange (DMX)
  - Australian War Memorial (AWM), Australian National Botanical Garden (ANBG)
- To build processing tool that can the and generate analysis of user purchasing and behaviors from Piction e-comm system
- Data analysis techniques, programming, database skills (relative to buzzword: Big data)
MOTIVATION

- Recommender System
  - Improving user experience, providing better product demonstration and then increasing sales conversion.
  - Identifying potential customer needs.
  - Amazon: 30% sales increasing
GOAL OF THE PROJECT

- Using data mining techniques to do customer behavior and purchasing analysis, and attempting to solve the following issues:
  - Recommendations for selling products better
  - Selling statistics and usage (most popular product)
  - Correlation analysis (key patterns in selling and usage)
DATA STRUCTURE

- Basket_status
- Basket_billing
- Basket_payment
- Basket_comments
- Basket_delivered
- Basket_pricing
- Basket_shipping
- Basket_totals
- Basket_contact
- Basket_grouped
- Basket_pricing
- Login_session
**DATA STRUCTURE**

- **purchase table**

```xml
<ROWSET>
  <ROW>
    <POID>628958</POID>
    <OPK>7042796</OPK>
    <DATE_CREATED>07-FEB-14</DATE_CREATED>
    <SESSID>327201707</SESSID>
    <P1D>2361</PID>
    <AID>265539</AID>
    <ACCTID>7028338</ACCTID>
    <ORDER_EXPIRED>F</ORDER_EXPIRED>
    <SCN>1</SCN>
    <BSK_STATUS_SCN>1</BSK_STATUS_SCN>
    <CUST_CTRL_RULE>6679443</CUST_CTRL_RULE>
    <CUST_EVENT_RULE>6679047</CUST_EVENT_RULE>
    <TAX_RULE>6677764</TAX_RULE>
    <PAYMENT_RULE>6679431</PAYMENT_RULE>
    <DEFAULT_PB>6677758</DEFAULT_PB>
    <STATUS_PROCESSING>Being Processed</STATUS_PROCESSING>
    <STATUS_LAB_SENT>Dispatched</STATUS_LAB_SENT>
    <WORKFLOW_CC_START>Order Payment Received</WORKFLOW_CC_START>
    <WORKFLOW_CC_DELIVERY>Order Payment Received</WORKFLOW_CC_DELIVERY>
    <WORKFLOW_CHEQUE_START>Order Payment Received</WORKFLOW_CHEQUE_START>
    <WORKFLOW_CHEQUE_DELIVERY>Order Payment Received</WORKFLOW_CHEQUE_DELIVERY>
    <WORKFLOW_NOPAY_START>Order Payment Received</WORKFLOW_NOPAY_START>
    <WORKFLOW_NOPAY_DELIVERY>Order Payment Received</WORKFLOW_NOPAY_DELIVERY>
  </ROW>

  <ROW>
    <POID>628957</POID>
    <OPK>7042795</OPK>
    <DATE_CREATED>07-FEB-14</DATE_CREATED>
    <SESSID>243510303</SESSID>
    <P1D>1</PID>
    <ACCTID>2361</ACCTID>
    <ORDER_EXPIRED>F</ORDER_EXPIRED>
    <SCN>1</SCN>
    <BSK_STATUS_SCN>1</BSK_STATUS_SCN>
  </ROW>
</ROWSET>
```
DATA STRUCTURE

- There are 136,203 sessions appeared in purchase table
  max  min  total
  6     1    158,867

- There are 23 kinds of status appeared in purchase table
  max  min  total
  58,497 1    58,867

- There are 6 kinds of comments appeared in purchase table
  max  min  total
  57,563 1    57,569
METHODODOLOGY/TECHNIQUES

- Similarity-based entity resolution (ER) techniques
- Community detection (CD)
  - Correlation (item to item, people to people)
EXPERIMENT

- Environment
  Java programming interface, Postgres database

- Dataset
  15 xml files

- Tasks
  - Convert xml into tables in Postgres DB
  - Analyze patterns based on customer purchase history, the purchase process (e.g., Pricing, delivery), customer search behavior, interaction between behaviors
CHALLENGES

- Data amount is huge
- Data is not clean
- Data is not completed