# Spark and Shark Bridges the Gap Between Business Intelligence and Machine Learning at Yahoo! Taiwan

Wisely Chen

# 2 2 One

# Agenda

- Who are we?
- What is the problem?
- What is the solution?
- Q&A

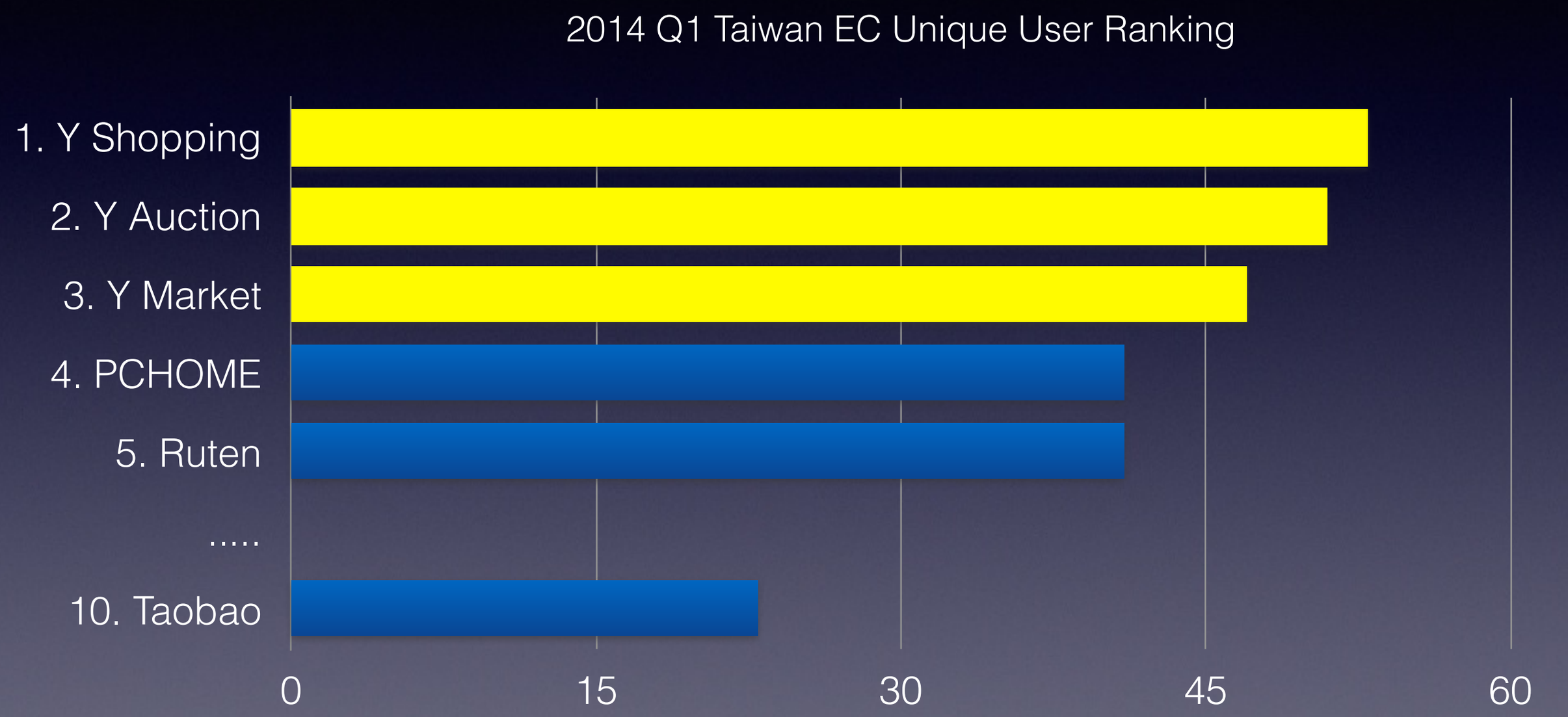
# Abbreviations

- EC : E-Commerce
- BI : Business Intelligence
- ML : Machine Learning
- MSTR : MicroStrategy

# Who I am

- Wisely Chen ( thegiive@gmail.com )
- Sr. Engineer in Yahoo[Taiwan] data team
- Loves to promote open source tech

# Who are we?




Data source: EagleEye 2014 Q1 report

# What is the problem?

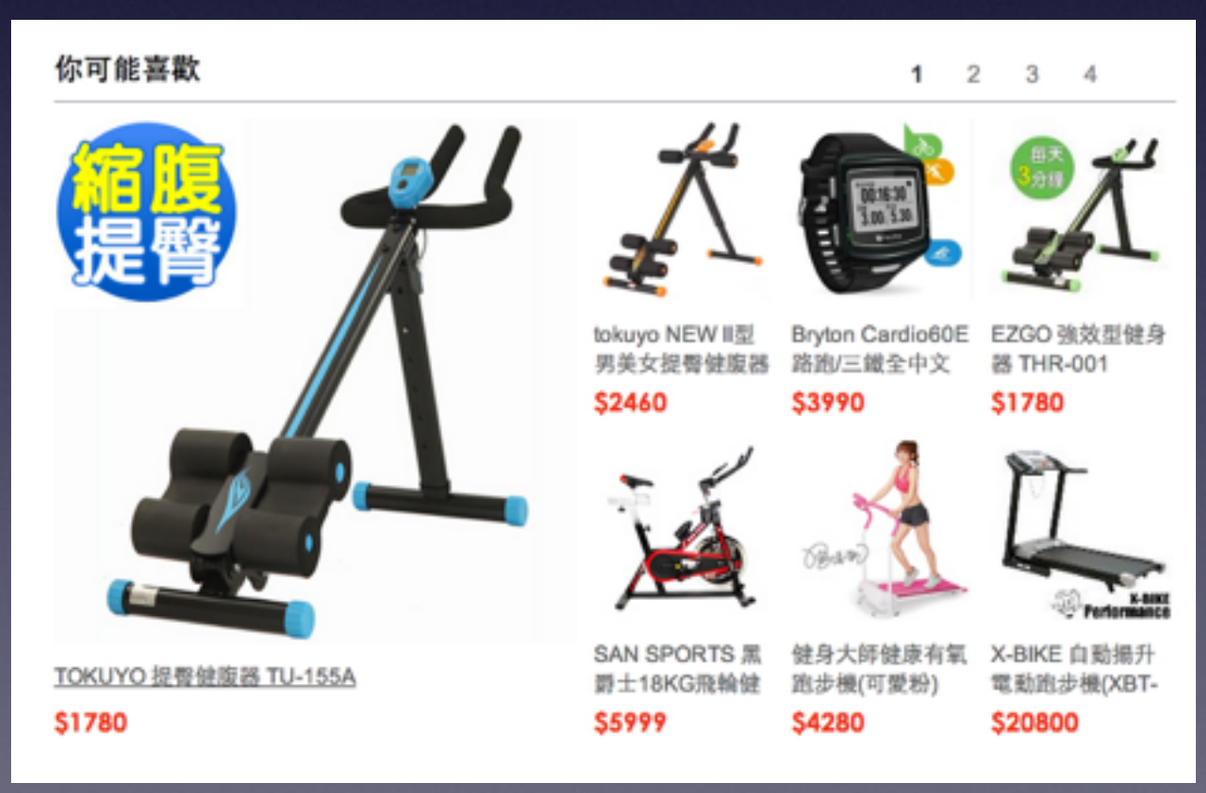
2 different products

2 types of data

# 2 different products

BI


ML



# 2 types of data


Traffic Data
User's views, clicks


Transaction Data
User's checkout

# Traffic Data

Beacon Service

Data Highway

Filtered Data (HDFS)

*Payment API*

*ERP (MySQL)*

*Data Mart (Oracle)*

# Transaction Data

# Traffic Data

Beacon Service

Data Highway

Filtered Data (HDFS)

*Payment API*

*ERP (MySQL)*

*Data Mart (Oracle)*

# Transaction Data

2 different products

*want to leverage*

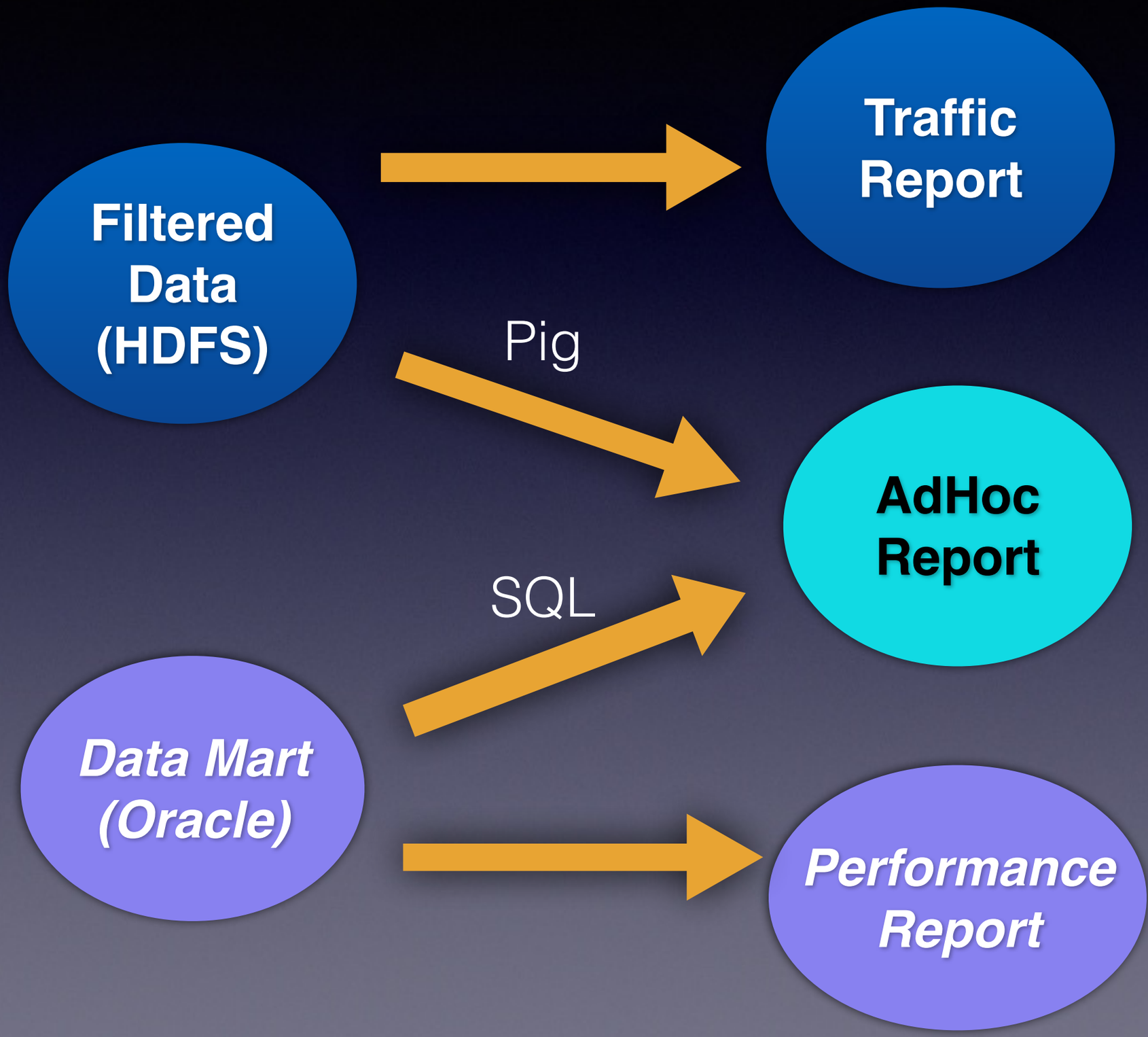
2 types of data

# Business Intelligence

BI

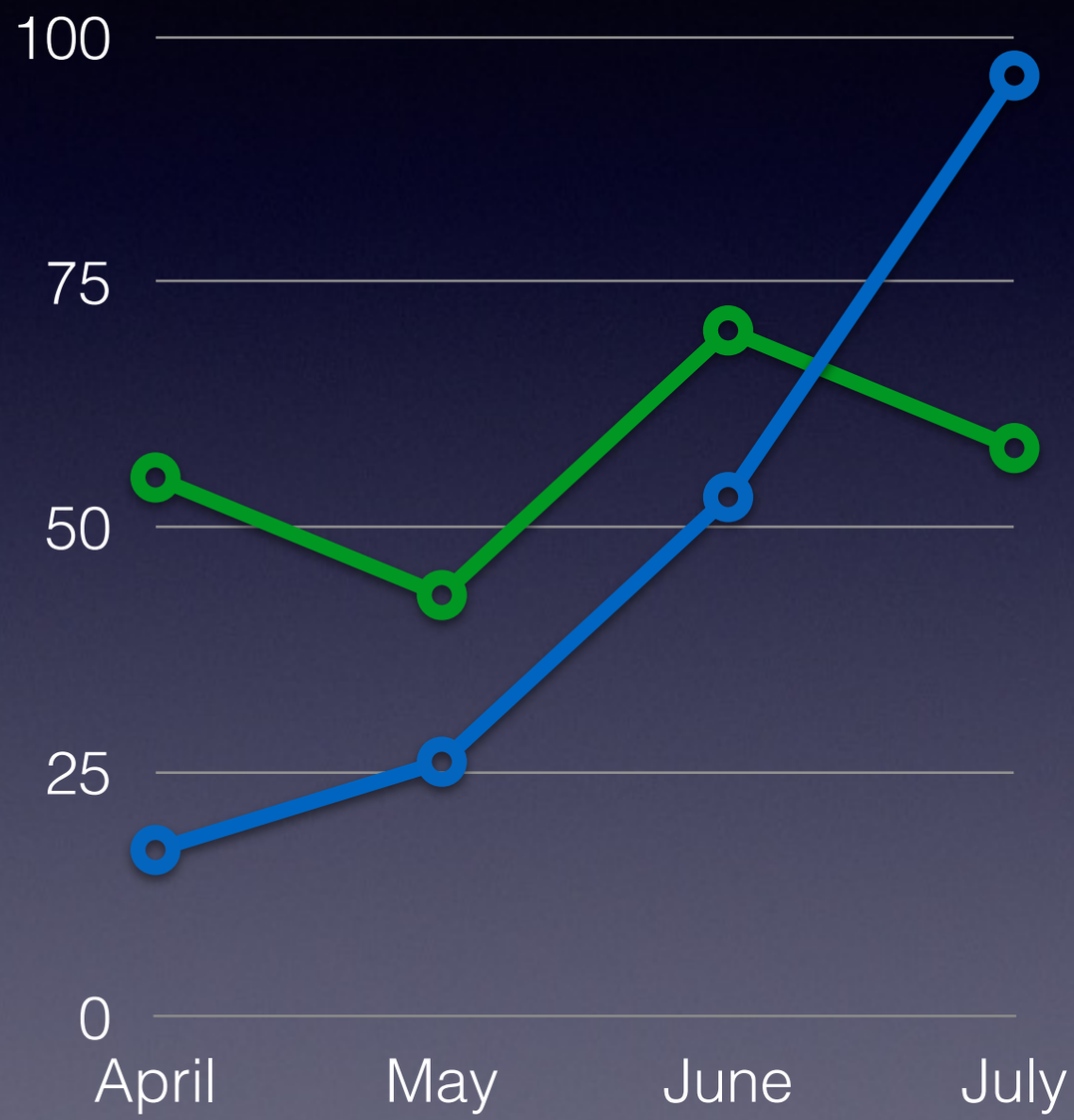

ML

# Original BI System

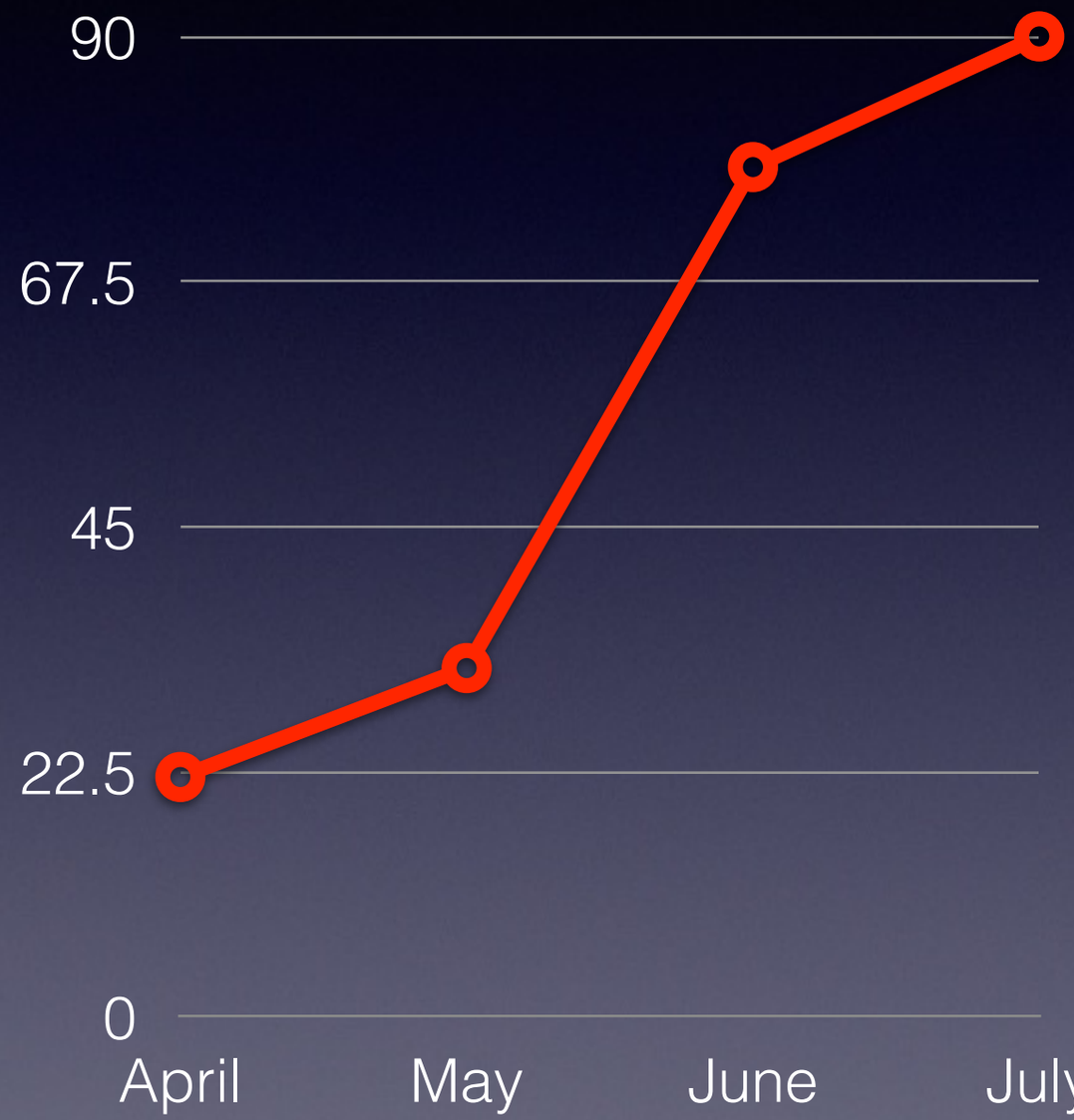


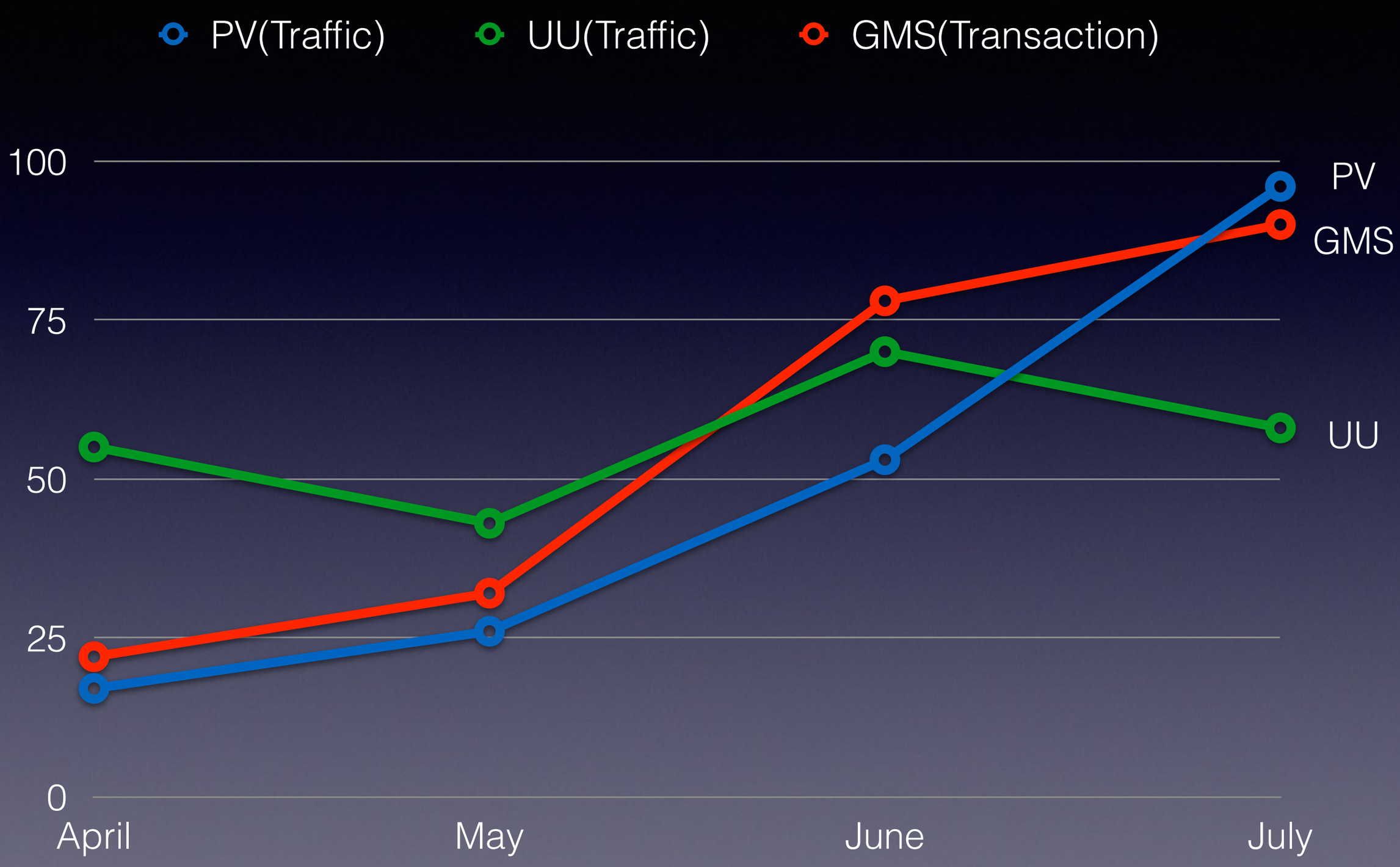
PV(Traffic)

UU(Traffic)



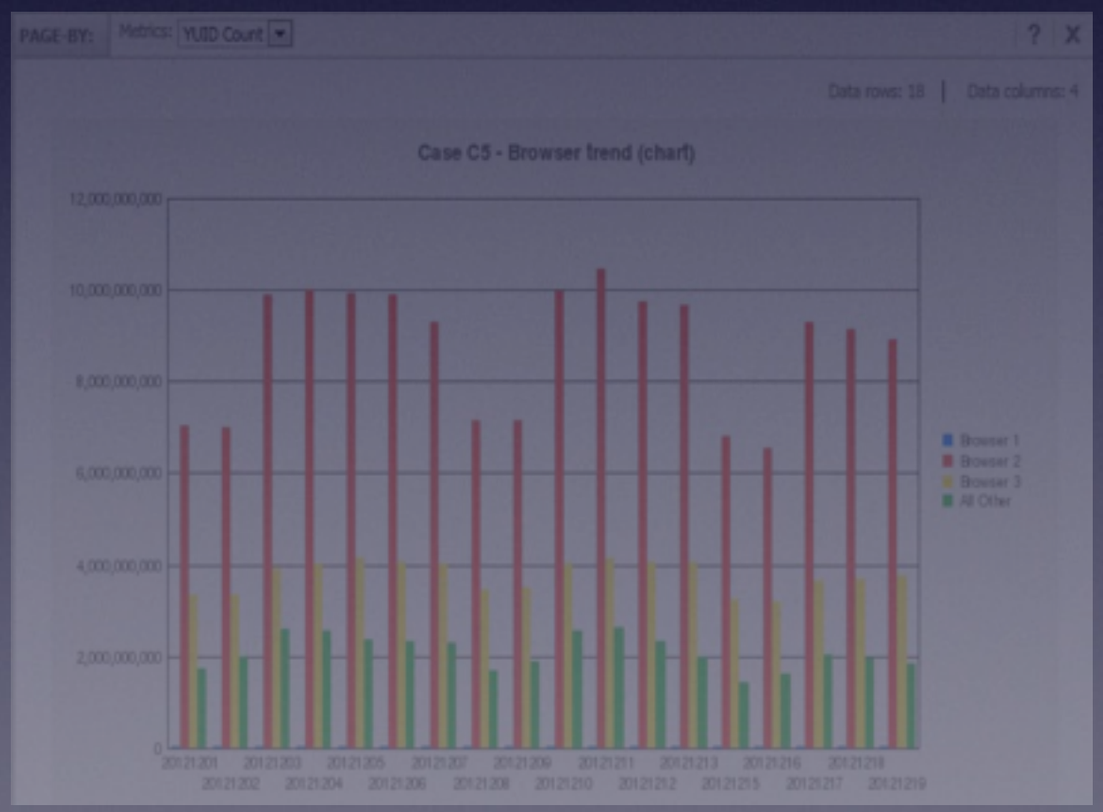

GMS(Transaction)



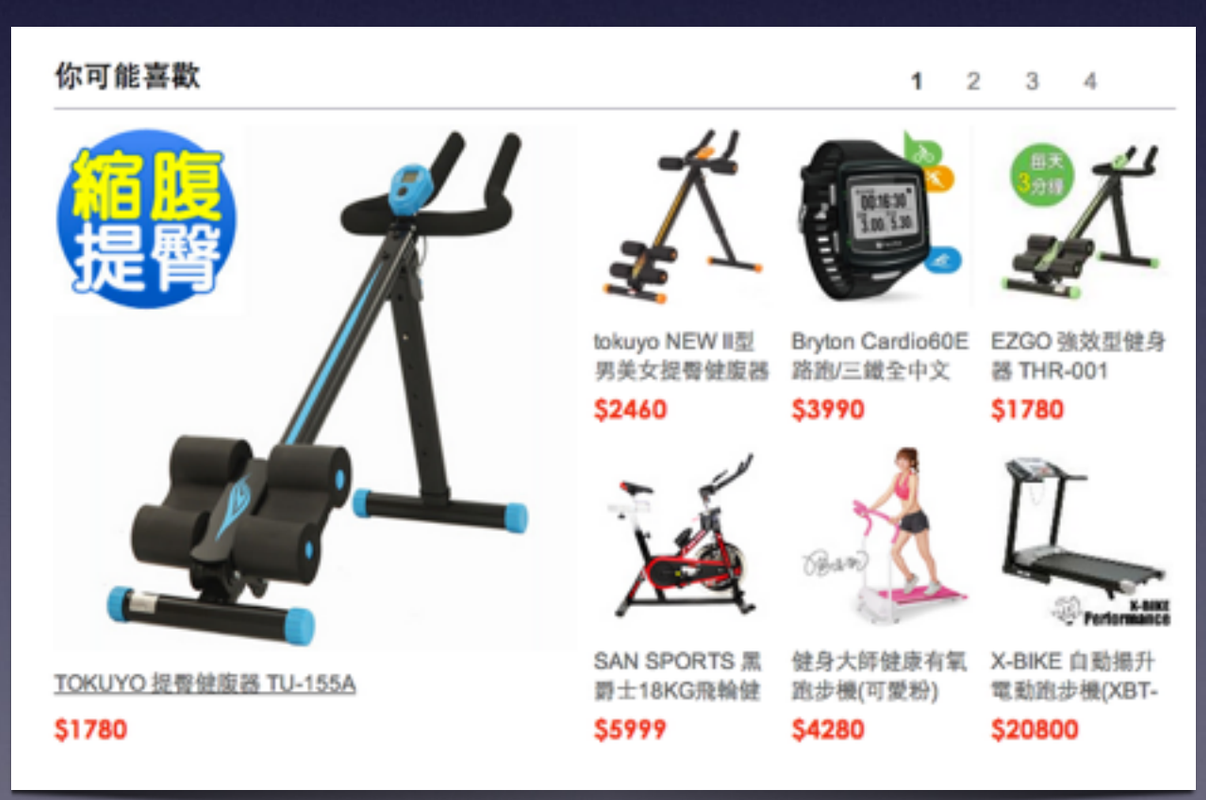

PV(Traffic)
UU(Traffic)
GMS(Transaction)
100
75
50
25
0
April
May
June
July
PV
GMS
UU

# Machine Learning

BI



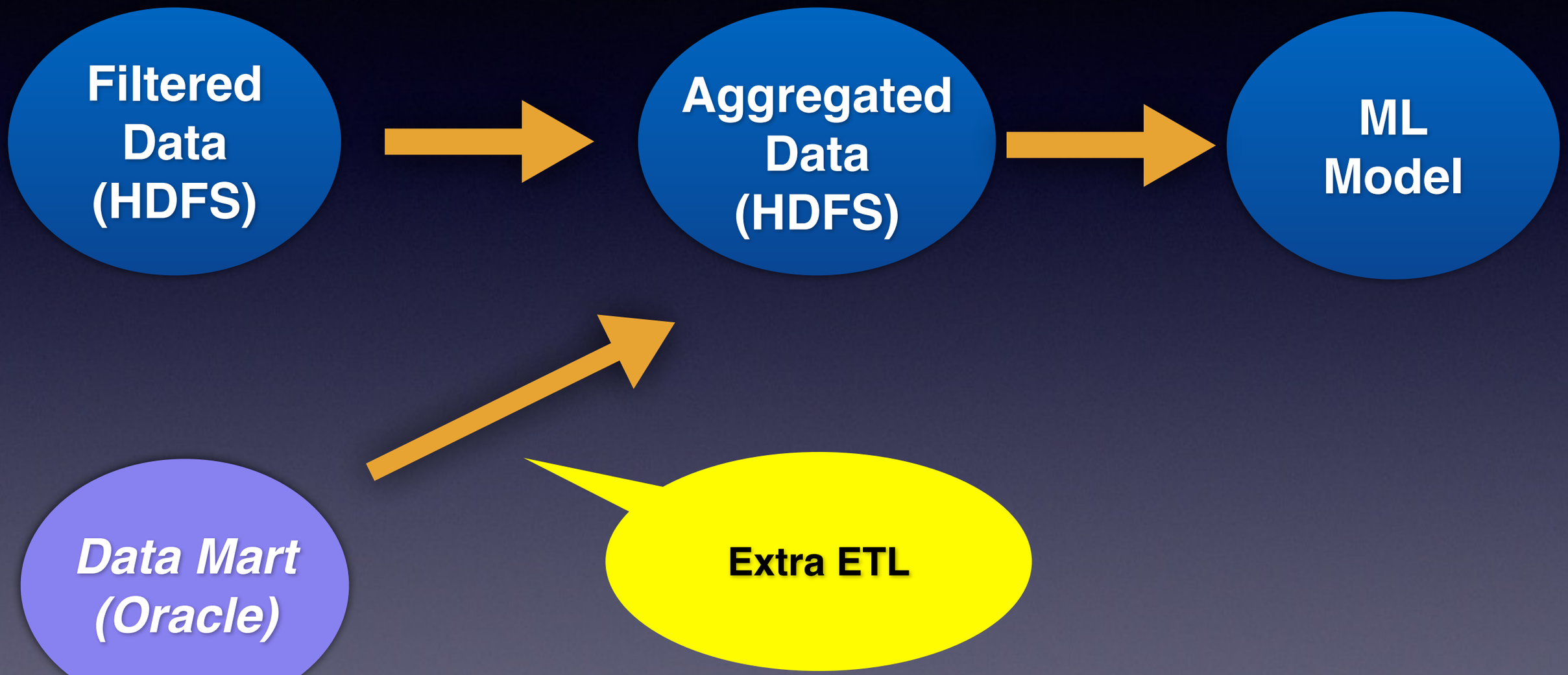

ML



# 2 types of data


Traffic Data
User's views / clicks
"Weak intention"


Transaction Data
User's checkout
"**Strong** Intention"


Filtered Data (HDFS)
Aggregated Data (HDFS)
ML Model
Data Mart (Oracle)
Extra ETL

2 different products

2 types of data

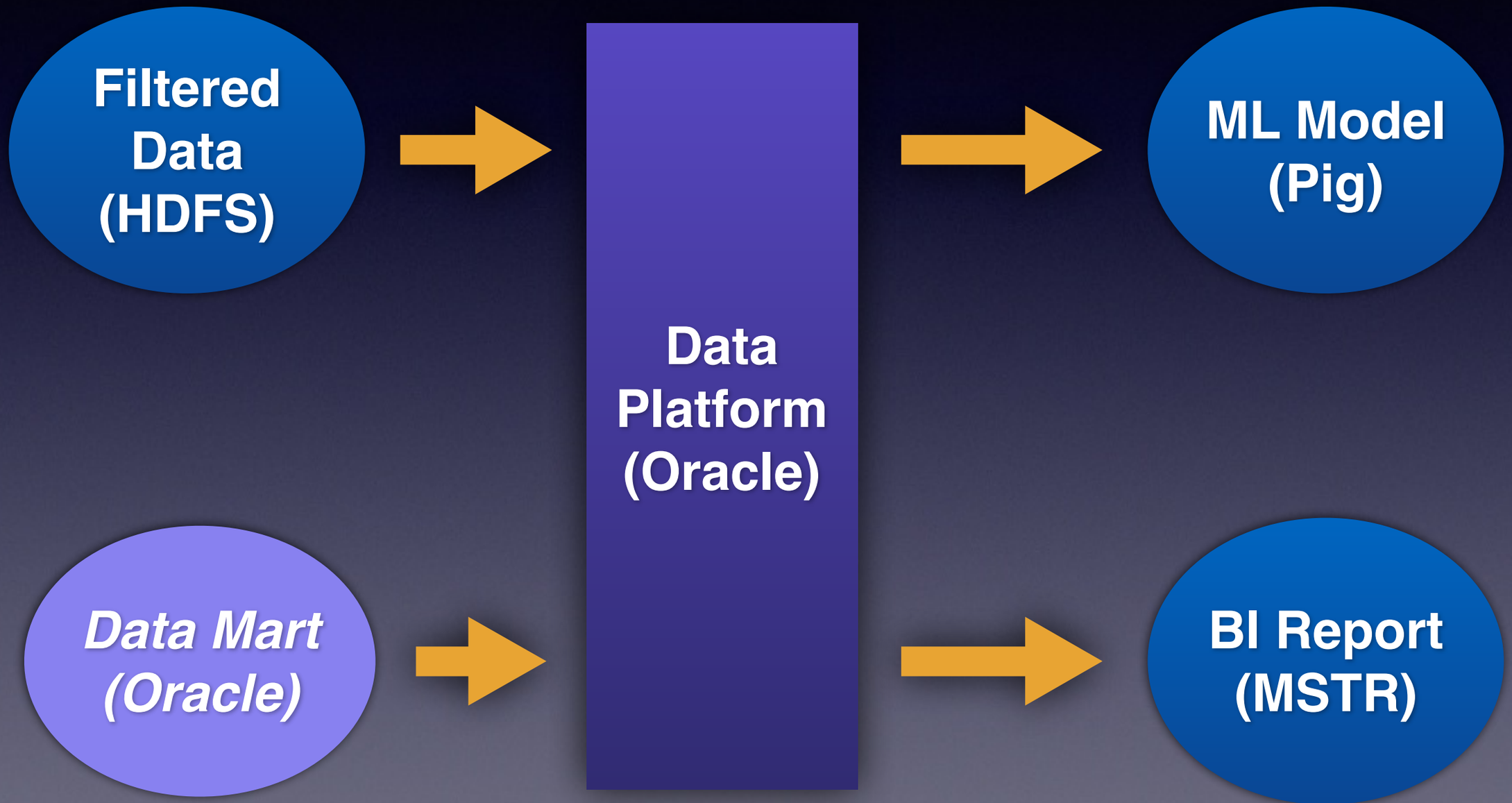
**One** unified platform

# What is the solution?

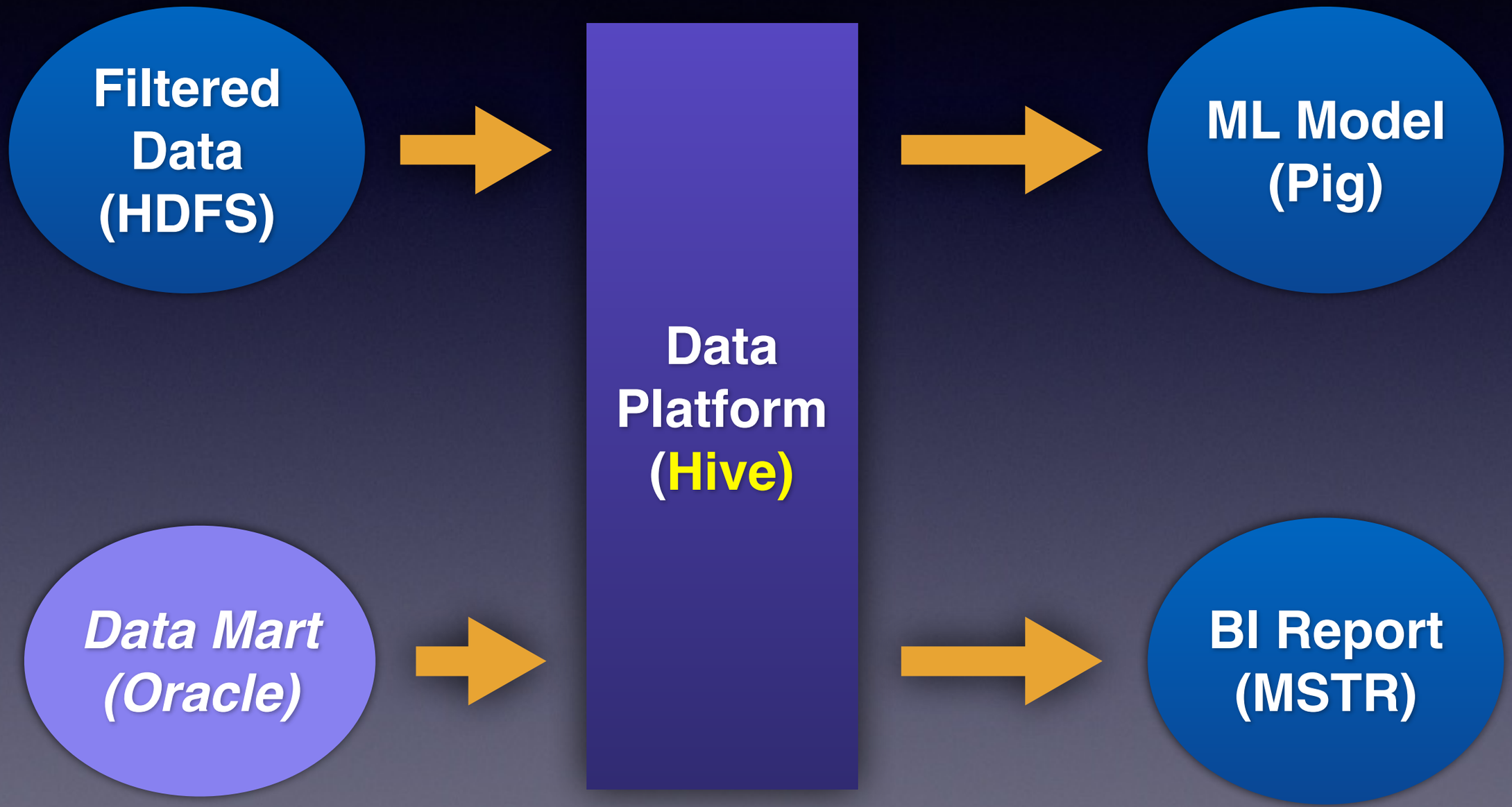
# Requirements

- Unified data platform for ML and BI
- BI tool (MicroStrategy) support ODBC
- Traffic data amount is greater than transaction data

# First Thoughts



# Hive as platform



# Hive is good

- Handles scale issues well
- Connects MicroStrategy with JDBC
- Integrates with ML language(pig)
- Speed is above average
    - Average response time is less than 20 sec (293T data in Hadoop 300+ nodes)

2 different products

2 types of data

**One** unified platform

We published this at the Hadoop Summit 2013

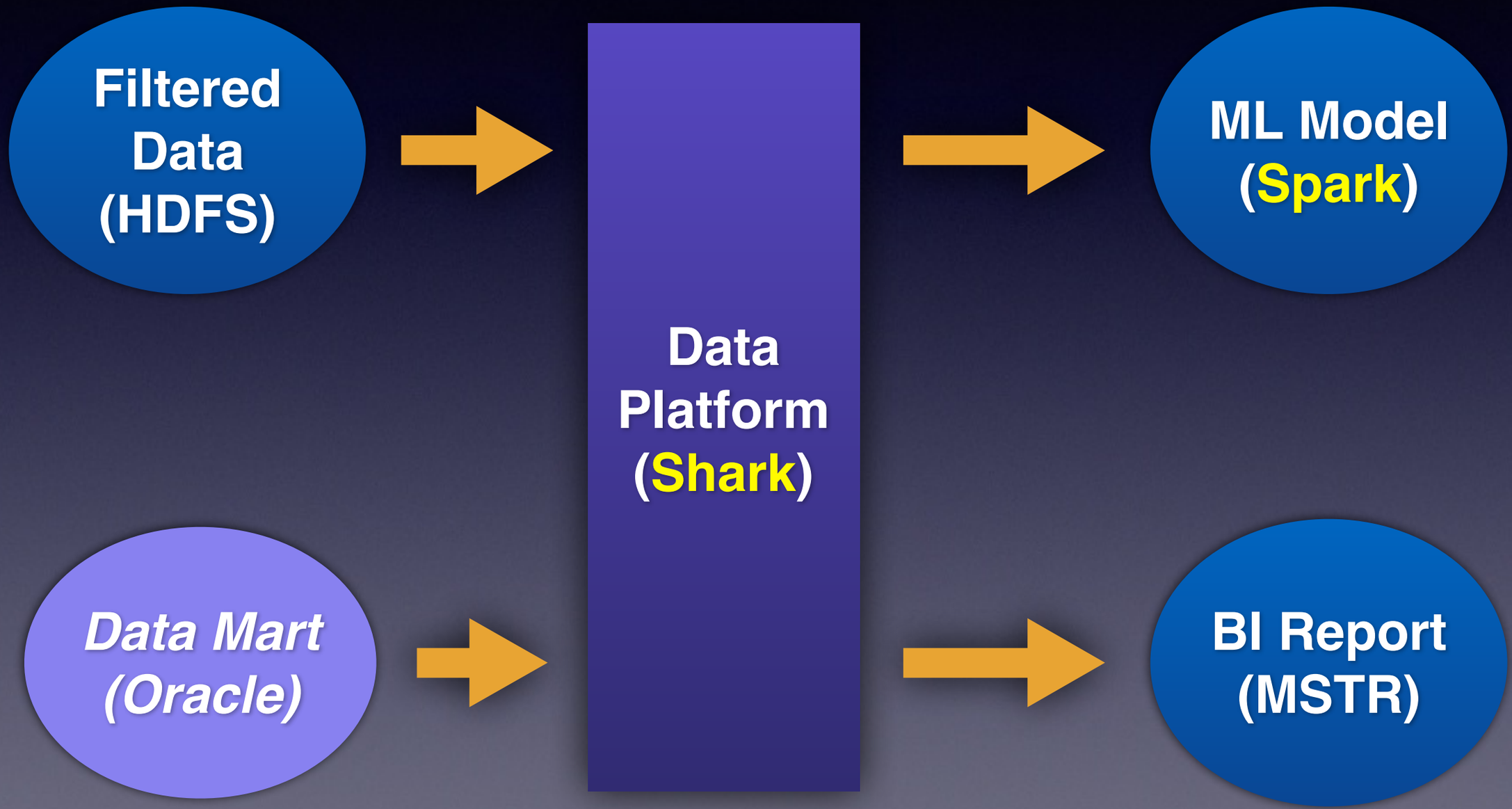

And we discovered Spark and Shark

2 different products

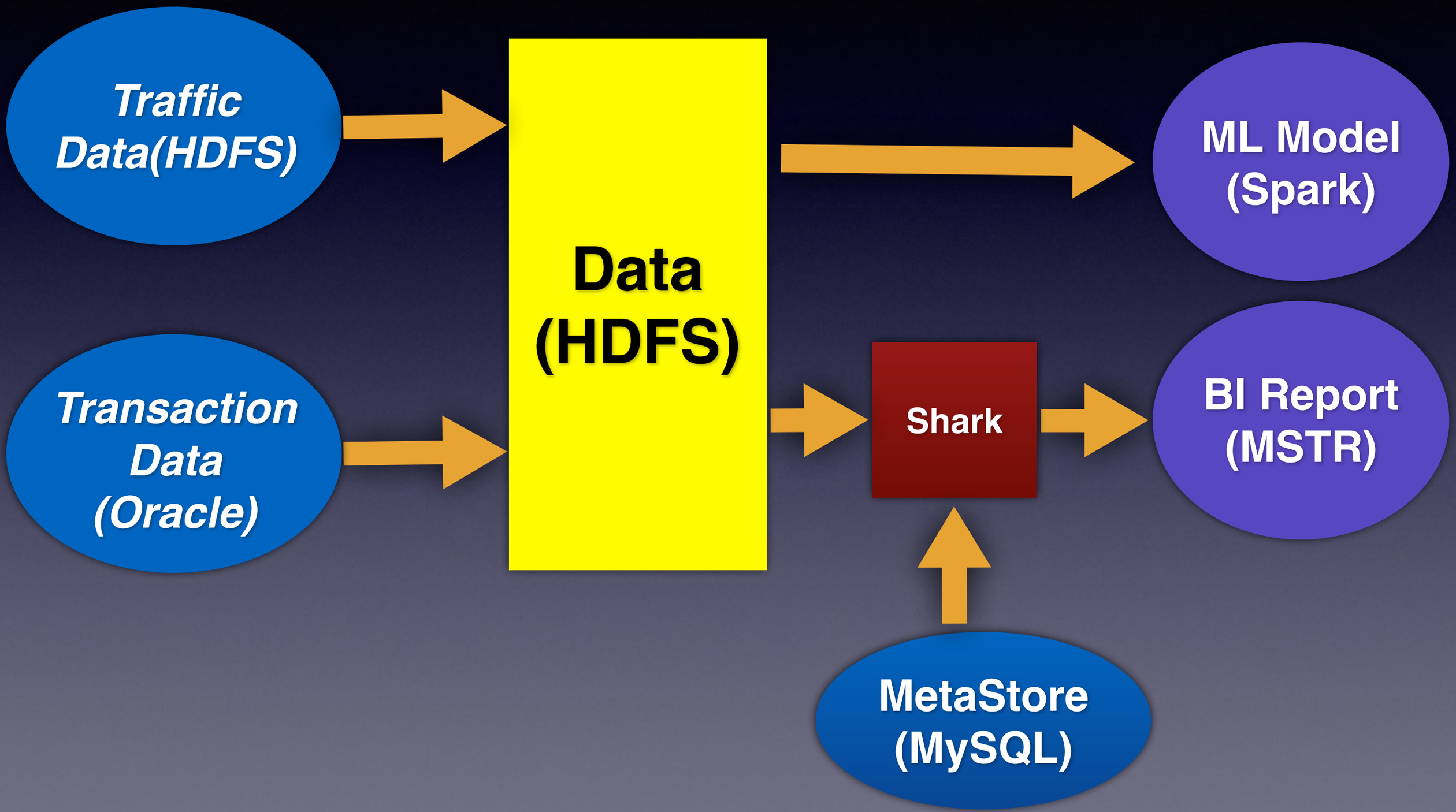
2 types of data

**One better** unified platform

# Spark / Shark to the rescue



# Architecture



# Why did we make this decision?

| | Hive | Shark |
|---|---|---|
| Scale | Y | Y |
| MSTR integration | ODBC | ODBC |
| Speed | OK | Much Better |
| ML Language Integration | Pig | Spark |

# Shark is like turbo-charged Hive

- Project migration cost from Hive is low
- BI tool integration
- Much better speed

# ML world moves to Spark

- Spark is significantly faster than MapReduce in most  machine learning cases
- Our ML team is migrating to Spark
- Spark and Shark can integrate better

# Important Config

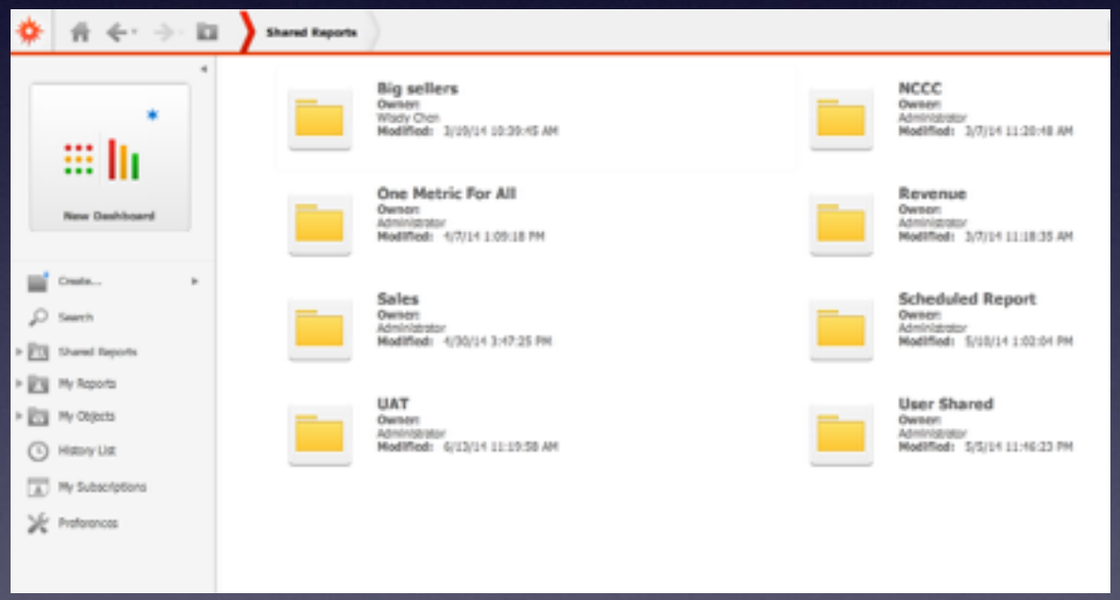
- Hardware config
  - Memory
  - Disk
  - Network interface
- Software Config
  - memoryFraction
  - spark.local.dir

# Lessons Learned

- Shark hangs sometimes (once per month)
- Join is time consuming in Shark / Hive
- Some queries are extremely slow in Shark 0.7
  - “ select count ( distinct xxx ) … group by…”

# As a result

- Launch in 2013 Q4
- 83% speed improvement in daily report
- 7.5 times faster in  ML model training



# As a result…

2 different products

2 types of data

**One better** unified platform

# Future work

- Evaluate SparkSQL
- Evaluate Catalyst

# Q&A