Building a Financial Data Warehouse: A lesson in empathy

Solmaz Shahalizadeh
Data Analysis Lead, Shopify
@solmaz_sh

@solmaz_sh

Data Team Lead @ Shopify

Statistics/ Machine Learning in Cancer Research

Analytics and development in Banking

what is this talk about?

1. How did our data journey start at Shopify?
2. Why we needed a massive change?
3. Change is difficult, how did we push forward?
4. After all of this, where are we now? What next?
5. Takeaways

shopify

Ways to sell Pricing Blog More

Log in

Shopify is everything you need to sell anywhere

Start your free 14-day trial today!

Email address

Get started

ONLINE STORE

- Overview
- Blog Posts
- Pages
- Themes
- Navigation
- Domains

Overview

Today

Sales

	Total	Orders
Today	$199.98	2
Yesterday	$119.97	3
Last 7 Days	$829.87	13
Last 30 Days	$3.9k	62
Last 90 Days	$10.3k	164

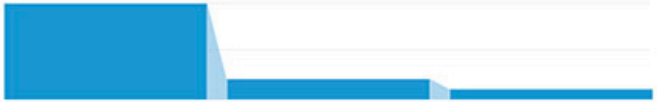
Visitors

	Total	Unique
Today	1337	1322
Yesterday	20k	17.5k
Last 7 Days	152k	105k
Last 30 Days	565.8k	358.9k
Last 90 Days	1.8m	992.6k

Conversions

Today

Added to Cart	Reached Checkout	Purchased
0.33% +0.33%	0.33% -0.5%	0% +0%



10 customers — 2 customers — 0 customers

Top Products

Today

Muscles Leggings	4 sold
Muscles Leggings	4 sold
Muscles Leggings	4 sold

Traffic Sources

Today

Direct	88% — 995 visitors
Referrals	7% — 74 visitors
Search Engines	5% — 60 visitors

Social Referrals

Today

Facebook	8%	10 visitors
Twitter	0%	0 visitors
Pinterest	0%	0 visitors

Top Countries

Today

United States	78%	190
Canada	7%	17
Russia	4%	10
United States	78%	190
Canada	7%	17
Russia	4%	10
Canada	7%	17
Russia	4%	10

Top Referrers

Today

m.facebook.com	54%	20
l.facebook.com	16%	6
www.facebook.com	11%	4
iris.josephnogucci.com	5%	2
josephnogucci.us5.list-manage.com	3%	1
www.facebook.com	3%	1
iris.josephnogucci.com	3%	1
josephnogucci.us5.list-manage.com	3%	1

Top Search Terms

Today

noqucci	25%	1
joseph nogucci	25%	1
josephnogucci	25%	1
prosperity buddha...	25%	1

About Shopify

Shopify is a leading cloud-based, multichannel commerce platform designed for small and medium-sized businesses. Merchants can use the software to design, set up and manage their stores across multiple sales channels, including web, mobile, social media, marketplaces, brick-and-mortar locations, and pop-up shops. The platform also provides a merchant with a powerful back-office and a single view of their business. The Shopify platform was engineered for reliability and scale, using enterprise-level technology made available to businesses of all sizes. Shopify currently powers over 200,000 businesses in approximately 150 countries, including: Tesla Motors, Budweiser, Red Bull, LA Lakers, the New York Stock Exchange, GoldieBlox, and many more.

Key Facts

SHOPIFY PLATFORM RELEASED
2006

MERCHANTS IN
~150 Countries

EMPLOYEES
900+

APPS IN OUR APP STORE
1000+

MERCHANTS
200,000+

THEMES IN OUR THEME STORE
100+

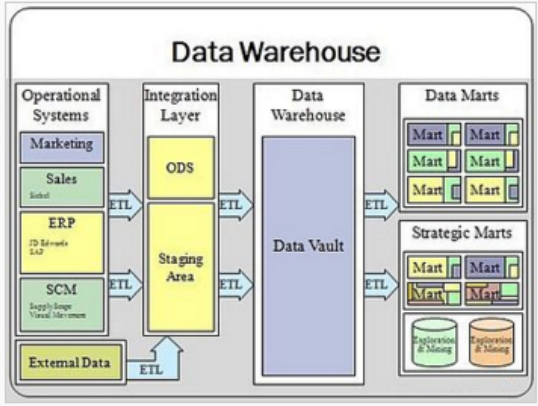
TOTAL SALES ON SHOPIFY
$12 Billion

EXPERTS IN OUR NETWORK
680+

Where did our data journey start?

What is a Data Warehouse?

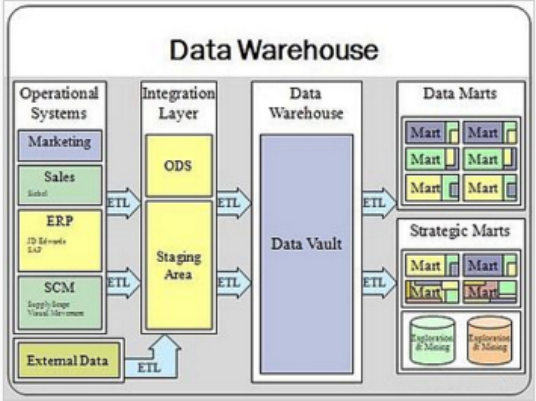
In computing, a **data warehouse** (**DW** or **DWH**), also known as an **enterprise data warehouse** (**EDW**), is a system used for reporting and data analysis. DWs are central repositories of integrated data from one or more disparate sources. They store current and historical data and are used for creating analytical reports for knowledge workers throughout the enterprise. Examples of reports could range from annual and quarterly comparisons and trends to detailed daily sales analyses.



Data Warehouse Overview

What is a Data Warehouse?

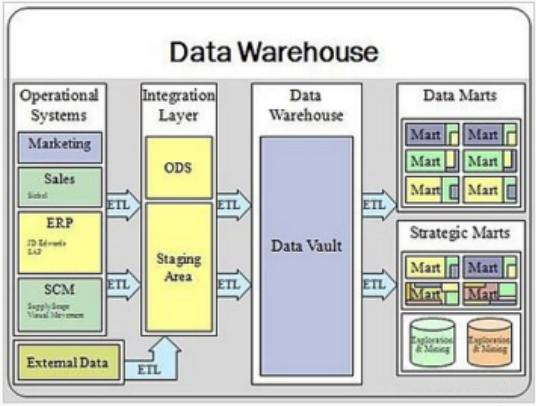
In computing, a **data warehouse** (**DW** or **DWH**), also known as an **enterprise data warehouse** (**EDW**), is a system used for reporting and data analysis. DWs are central repositories of integrated data from one or more disparate sources. They store current and historical data and are used for creating analytical reports for knowledge workers throughout the enterprise. Examples of reports could range from annual and quarterly comparisons and trends to detailed daily sales analyses.



Data Warehouse Overview

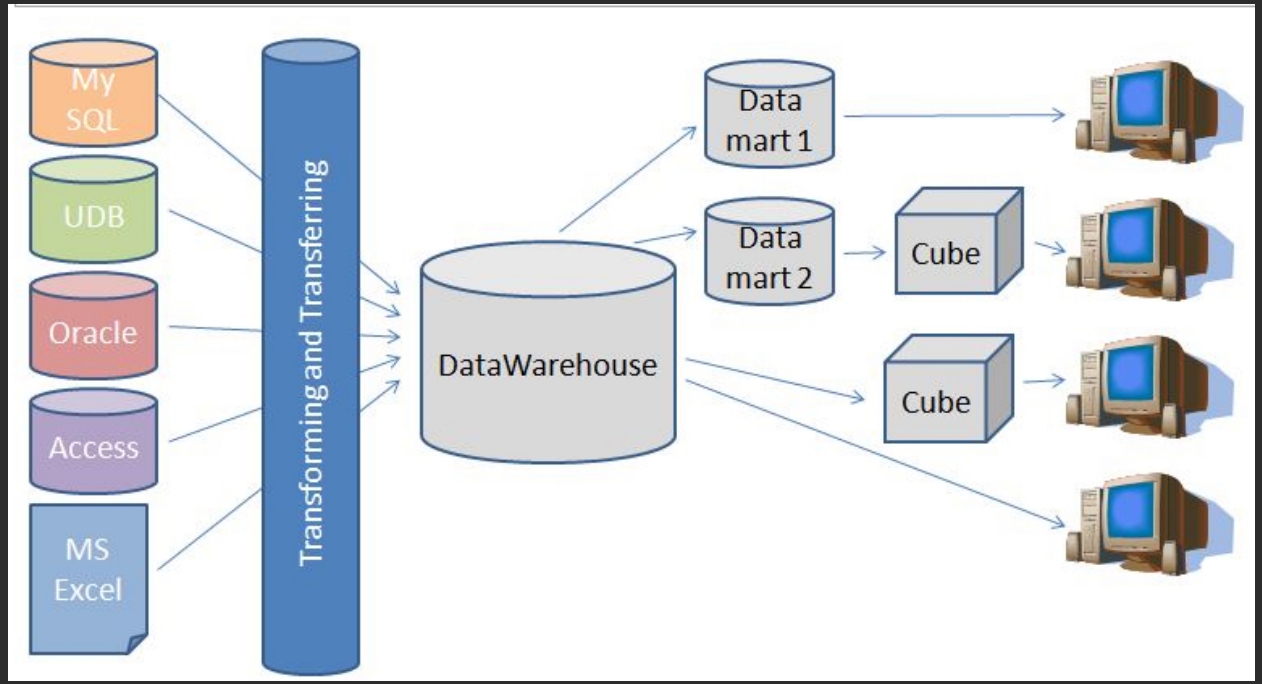
What is a Data Warehouse?

In computing, a **data warehouse** (**DW** or **DWH**), also known as an **enterprise data warehouse** (**EDW**), is a system used for reporting and data analysis. DWs are central repositories of integrated data from one or more disparate sources. They store current and historical data and are used for creating analytical reports for knowledge workers throughout the enterprise. Examples of reports could range from annual and quarterly comparisons and trends to detailed daily sales analyses.



Data Warehouse Overview

ETL systems overview



1st generation data warehouse

ETL:

- E: xtract from Production MySQL and other sources using in-house ruby etl
- !T: very little (almost no) T in the ETL
- L: oad into a Vertica database

* a copy of “any” operational table in the warehouse

1st generation reporting

In-house reporting (SQL and coffeescript):

- Clean the data at the report level using SQL
- Define the metric in the report
- use D3 for visualizations

* JIT metric definition

KPI for an e-commerce SaaS company

Customers

GMV (Gross Merchandise Volume)

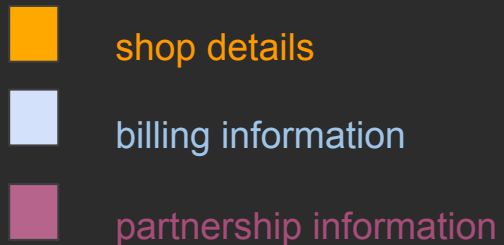
Revenue

KPI for an e-commerce SaaS company

Customers: *by cohort, by monthly recurring revenue, by partnership*

GMV: *by payment gateway, by industry, by monthly recurring revenue*

Revenue: *by cohort, by subscription type, by partnership*




hard to connect silos
image link https://goo.gl/aOkbXU

Different business areas, different definitions

- What is the geographical location of a customer?
- What do we mean when we talk about cohorts?
- Did the customer churn?
- ...

Things that are called the same thing, do not mean the same thing

image link https://goo.gl/VJwFYx

Exploratory analysis for all:

While analysts idea of question!= business idea of question:

1) write ad-hoc SQL query
2) assume analyst 100% understands the business process
3) assume the ad-hoc cleaning and conforming of values is reproducible
4) extract data and sending data dump
5) upon seeing the results, business user realizes this is not what they wanted


sad users & analysts
image link https://goo.gl/zdatv6

“The world as we have created it is a process of our thinking. It cannot be changed without changing our thinking.”

- Albert Einstein

Dimensional modelling

Dimensional Modelling is a design technique for databases intended to support end-user queries in a data warehouse. It is oriented around understandability and performance.

- Wikipedia

Focus on a business process:

- What are the dimensions?
- What are the measures?
- What is the grain?

Focus on modelling a process

Not a specific source tables but a business process

e.g: processing a financial transaction in exchange for a product as part of GMV

Define dimensions, facts, grain

Business Process: processing financial transaction

Dimensions: *payment gateway that processed the transaction, Industry of the shop, monthly recurring revenue (subscription amount) of the shop*

Measures: $ amount of transaction

Define dimensions, facts, grain

Business Process: processing financial transaction

Dimensions: *payment gateway that processed the transaction, Industry of the shop, monthly recurring revenue (subscription amount) of the shop*

Measures: $ amount of transaction

Remove anything without analytical value: name of the customer, exact address of the customer, etc.

image link https://goo.gl/b8jYVk

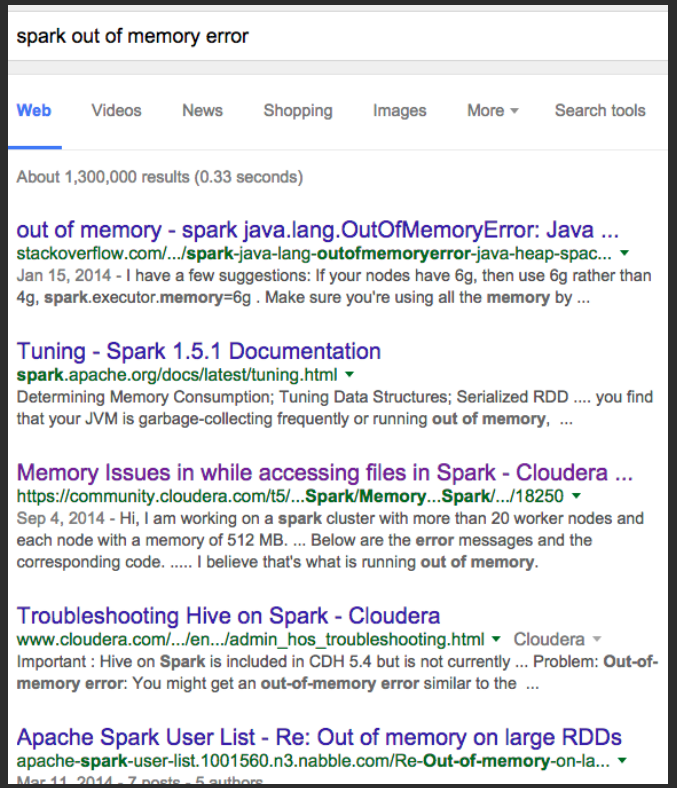
2nd generation data warehouse

E: in-house Ruby extracts to

T: Heavy transforms in

L: Load into Amazon Redshift

Building the ETL framework in Spark




```
org.apache.spark.SparkException: Job aborted due to stage failure: Task 0.0:0 failed 4 ti
Driver stacktrace:
        at org.apache.spark.scheduler.DAGScheduler.org$apache$spark$scheduler$DAGSchedule
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage$1.apply(DAGSchedul
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$abortStage$1.apply(DAGSchedul
        at scala.collection.mutable.ResizableArray$class.foreach(ResizableArray.scala:59)
        at scala.collection.mutable.ArrayBuffer.foreach(ArrayBuffer.scala:47)
        at org.apache.spark.scheduler.DAGScheduler.abortStage(DAGScheduler.scala:1015)
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$handleTaskSetFailed$1.apply(D
        at org.apache.spark.scheduler.DAGScheduler$$anonfun$handleTaskSetFailed$1.apply(D
        at scala.Option.foreach(Option.scala:236)
        at org.apache.spark.scheduler.DAGScheduler.handleTaskSetFailed(DAGScheduler.scala
        at org.apache.spark.scheduler.DAGSchedulerEventProcessActor$$anonfun$receive$2.ap
        at akka.actor.ActorCell.receiveMessage(ActorCell.scala:498)
        at akka.actor.ActorCell.invoke(ActorCell.scala:456)
        at akka.dispatch.Mailbox.processMailbox(Mailbox.scala:237)
        at akka.dispatch.Mailbox.run(Mailbox.scala:219)
        at akka.dispatch.ForkJoinExecutorConfigurator$AkkaForkJoinTask.exec(AbstractDispa
        at scala.concurrent.forkjoin.ForkJoinTask.doExec(ForkJoinTask.java:260)
        at scala.concurrent.forkjoin.ForkJoinPool$WorkQueue.runTask(ForkJoinPool.java:133
        at scala.concurrent.forkjoin.ForkJoinPool.runWorker(ForkJoinPool.java:1979)
        at scala.concurrent.forkjoin.ForkJoinWorkerThread.run
```

The Little Warehouse That Couldn't, Or: How We Learned to Stop Worrying and Move to Spark

Yandu Oppacher (Shopify)

Monday, June 15
5:45 PM – 6:00 PM
Imperial Ballroom (Level 2)

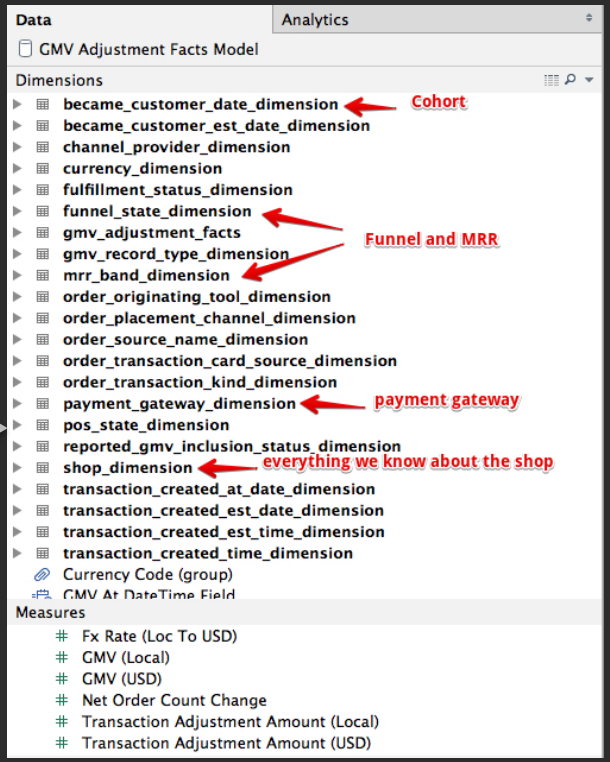
SLIDES PDF VIDEO

Shopify's commerce platform now powers over 150K merchants and continues to grow. The huge volume and variety of our data pushed our homegrown reporting and warehousing system to the edge. As the maintenance and performance costs became too much we moved to HDFS and Spark, using both python and scala to transform our vast amounts of operational data into dimensional models to provide better insight. Find out the lessons we've learned moving our entire organization onto fully conformed facts and dimensions, the clusters we've cratered, the walls we've hit, and what we did to overcome them to build our Starscream framework.

ABOUT YANDU

Yandu spent seven years working on the relational engine of IBM's Cognos BI and reporting tool, working with dozens of vendors and helping to rebuild their 15 year old engine. For the past year and a half he's been working on Shopify's data engineering team to help build a new framework for dimensionally modeling Shopify's data.

2nd generation reporting

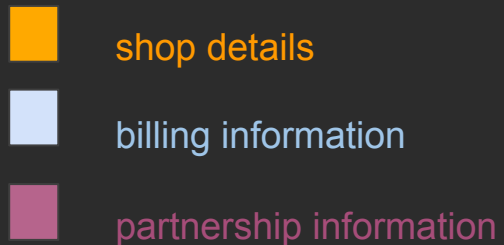


KPI for an e-commerce SaaS company

Customers: *by cohort, by monthly recurring revenue, by partnership*

Sales: *by payment gateway, by industry, by monthly recurring revenue*

Revenue: *by cohort, by subscription type, by partnership*



A walk in the data garden

Insightful reports

Self-service analytics (Tableau/ ipython)

New Business Analyst role

Faster feature selection (ML)

image link https://goo.gl/MtVWcJ

Takeaways

- Separation of business process/data model from ETL implementation
- Dimensional modelling for business process/ data models
- Clear data definitions to enable self service analytics
- Reproducible data models for reliable (and faster) predictive analysis

Questions?

We are hiring!

http://shopify.com/careers

Thank you

@solmaz_sh