



# US Mortgage Loan Data Analysis

Alyssa, Chris, Gerard, Seung

## Problem & Motivation

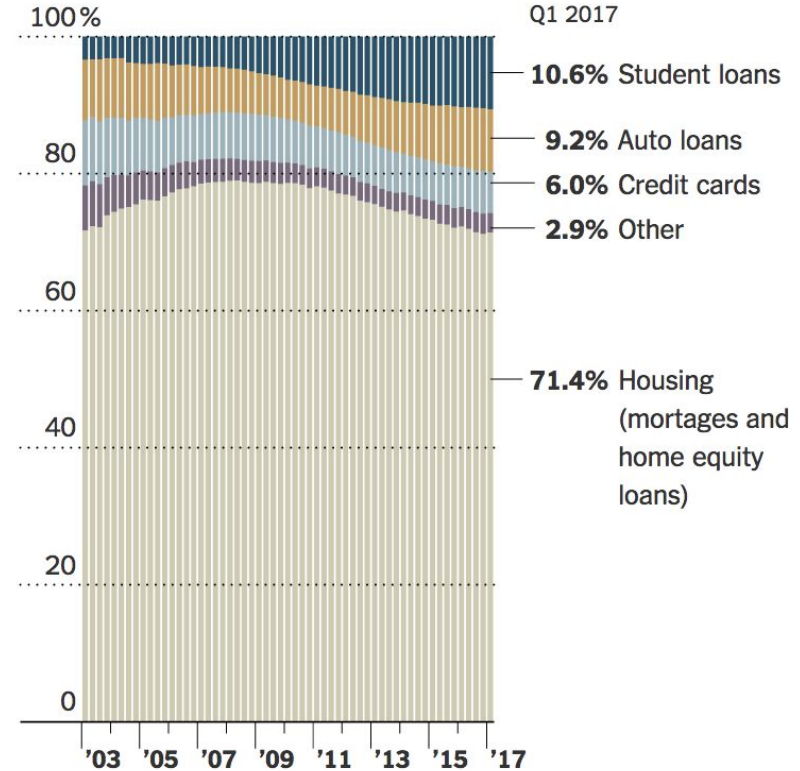
**Housing mortgages:** primary component of ~\$13 billion consumer debt in the United States.

Basis for an important class of financial asset; residential mortgage-backed securities (RMBS).

Predicting their future performance is a key task for securities analysts, especially the major ratings agencies.

STANDARD & POOR'S **MOODY'S**  
**Fitch** Ratings

SHARE OF CONSUMER DEBT



# Architecture Design

Acquire, store, and transform raw data: use Hadoop File System and Hive



**Download  
the data**



**Load  
into  
HDFS**



**Load  
into  
Hive**



**Transform  
and clean  
data**

Serving layer: Query Hive metadata tables and Tableau



**Allow open-ended querying  
of entire dataset through  
tool of users' choice**



**Create visualizations in  
Tableau dashboards**

# Data Acquisition and Organization:



Shell scripts  
for initiating  
environment

Most functions from initiating the system to transforming the data can be run by a series of .sh scripts



Download data  
using .py script

For loan or acquisition quarter, python script lets user download Fannie and Freddie data of interest



Shell script  
to load  
data into  
HDFS

Major loading and transformation is accomplished by .sh scripts, which can be part of fully automated download-load-transform process



Sql scripts for  
transforming  
data

System also allows for free roaming and querying of data beyond the pre-made tools.

# Data Acquisition and Organization:

Large amount of Data from each organization (0.5-10 GB based on time span)

Fannie and Freddie **not** entirely overlapping

Useful metrics, such as **Interest rate**, **debt-to-income**, **loan ID** match as keys

Other metrics, such as **maturity dates**, **delinquency status** are scaled differently

## Acquisition Data

Fannie Mae Acq		Freddie Mac Acq	
LOAN_ID	string	fico	string
ORIG_CHN	string	dt_first_pi	string
SELLER_NAME	string	flag_fthb	string
ORIG_RT	string	dt_mtr	string
ORIG_AMT	string	cd_msa	string
ORIG_TRM	string	mi_pct	string
ORIG_DTE	string	cnt_units	string
FRST_DTE	string	occpy_sts	string
OLTV	string	cltv	string
OCLTV	string	diti	string
NUM_BO	string	orig_upb	string
DTI	string	ltv	string
CSCORE_B	string	int_rt	string
FTHB_FLG	string	channel	string
PURPOSE	string	ppmt_pnlty	string
PROP_TYP	string	prod_type	string
NUM_UNIT	string	st	string
OCC_STAT	string	prop_type	string
STATE	string	zipcode	string
ZIP_3	string	id_loan	string
MI_PCT	string	loan_purpose	string
PRODUCT_TYPE	string	orig_loan_term	string
CSCORE_C	string	seller_name	string
ML_TYPE	string	servicer_name	string
RELOCATION_FLG	string	super_conform_flag	string
		pre_harp_loan_num	string

## Performance Data

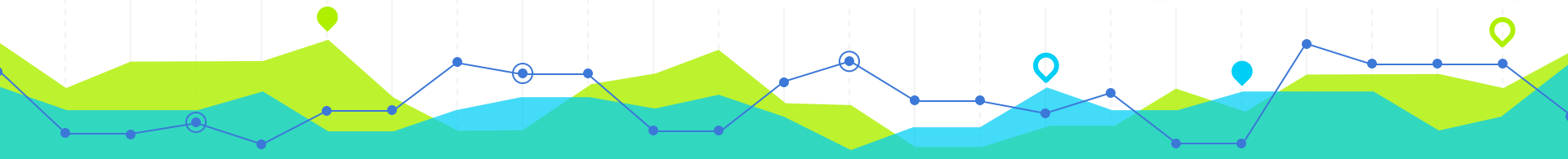
Fannie Mae Perf		Freddie Mac Perf	
LOAN_ID	string	ID_loan	string
MONTHLY_RPT_PRD	string	Period	string
SERVICER_NAME	string	Act_endg_upb	string
LAST_RT	string	delq_sts	string
LAST_UPB	string	loan_age	string
LOAN_AGE	string	mths_remg	string
MONS_TO_LEGAL_MAT	string	flag_mod	string
ADJ_MONTH_TO_MAT	string	CD_Zero_BAL	string
MATURITY_DATE	string	Dt_zero_BAL	string
MSA	string	New_Int_rt	string
DELQ_STATUS	string	Amt_Non_Int_Bmg_Upb	string
MOD_FLAG	string	Dt_Lst_Pi	string
ZERO_BAL_CODE	string	MI_Recoveries	string
ZB_DTE	string	Net_Sale_Proceeds	string
LPI_DTE	string	Non_MI_Recoveries	string
FCC_DTE	string	Expenses	string
DISP_DT	string	legal_costs	string
FCC_COST	string	maint_pres_costs	string
PP_COST	string	taxes_ins_costs	string
AR_COST	string	mics_costs	string
IE_COST	string	actual_loss	string
TAX_COST	string	modcost	string
NS_PROCS	string		
CE_PROCS	string		
RMW_PROCS	string		
O_PROCS	string		
NON_INT_UPB	string		
PRIN_FORG_UPB_FHFA	string		
REPCH_FLAG	string		
PRIN_FORG_UPB_OTH	string		
TRANSFER_FLAG	string		

# Data Acquisition and Organization:

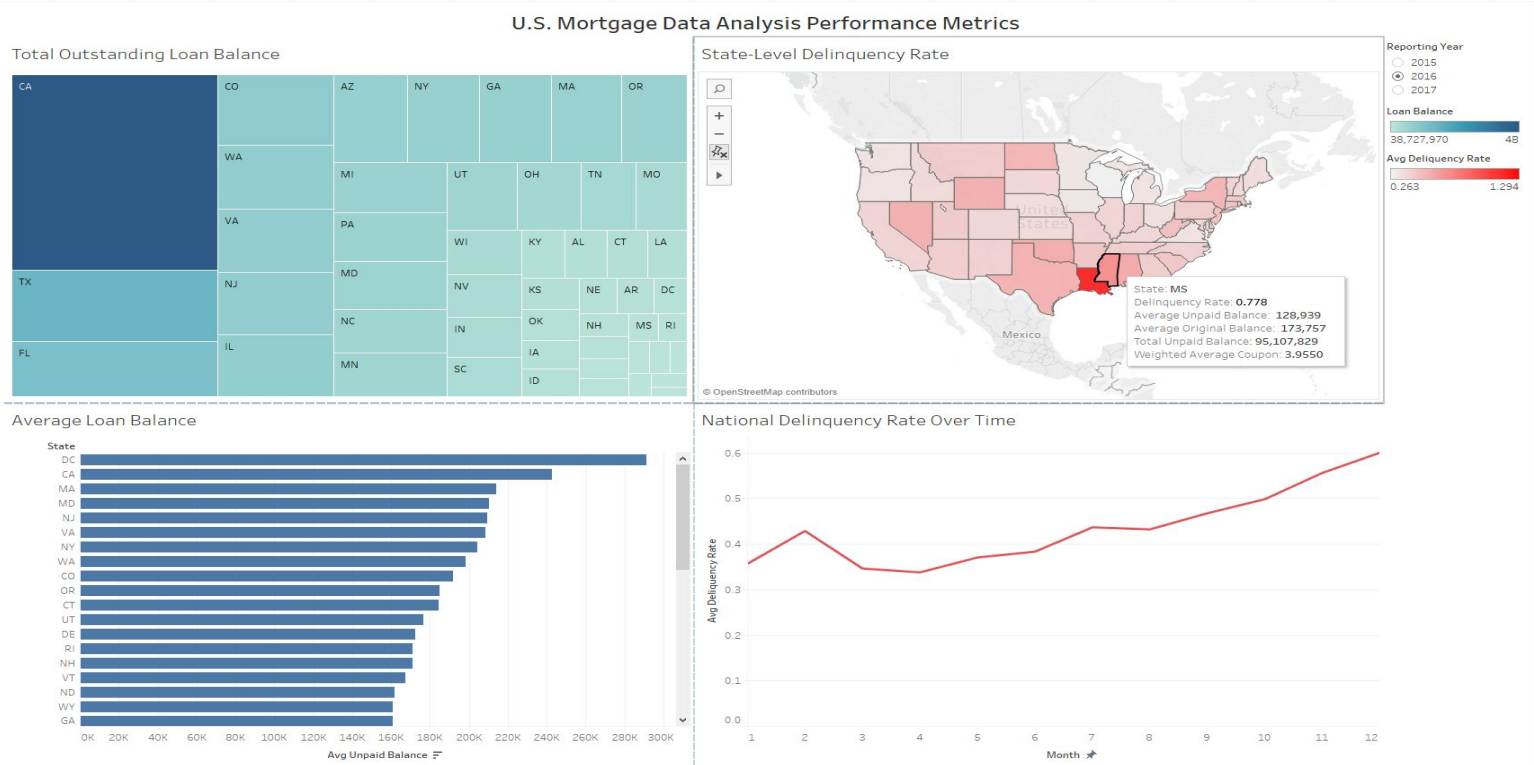
- Fannie and Freddie Acquisition & Performance data is joined
  - Depending on the time range, this could take many hours
  - Large data uploads and joins are required 4 times per year and can be performed overnight
- Data is then aggregated on loan origination state and 3 digit zip code (ex. 10036 = 100\*\*= 100)
- While user has access to all data, the aggregated State and Zipcode tables are especially useful for visualizations

3 Digit Zipcode Acquisition	
ZIP_3	string
FRST_DTE_MTH	string
FRST_DTE_YR	string
NUM_LOANS	string
SUM_ORIG_AMT	string
AVG_ORIG_AMT	string
AVG_ORIG_RT	string
WAVG_ORIG_RT	string
AVG_OLTV	string
WAVG_OLTV	string
AVG_DTI	string
WAVG_DTI	string
AVG_CSCORE	string
WAVG_CSCORE	string

3 Digit Zipcode Performance	
ZIP_3	string
FRST_DTE_MTH	string
FRST_DTE_YR	string
RPT_PRD_MTH	string
RPT_PRD_YR	string
NUM_LOANS	string
SUM_ORIG_AMT	string
SUM_UPB_AMT	string
AVG_UPB_RT	string
AVG_LAST_RT	string
WORG_AVG_LAST_RT	string
WUPB_AVG_LAST_RT	string
AVG_DELQ_RT	string
WORG_AVG_DELQ_RT	string
WUPB_AVG_DELQ_RT	string



# Data Serving: Tableau Example - Performance



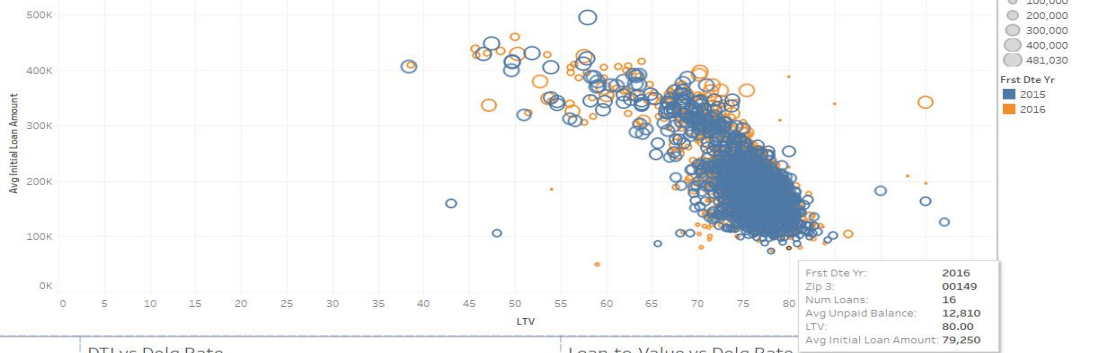
# Data Serving: Tableau Example

US Mortgage Data Analysis Origination Metrics

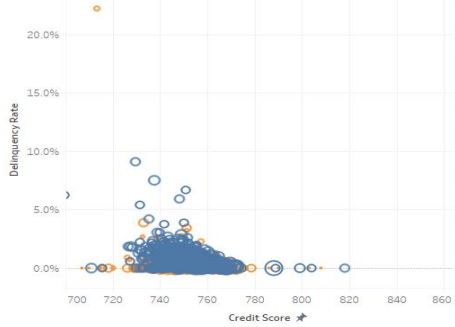
Credit Score vs Delq Rate

	Frst Dte Yr	
	2015	2016
Num Loans	2,652,273	2,386,232
Sum Orig Amt	\$94,749,508,000	\$55,247,826,000
Avg Initial Loan Amount	224,241	232,688
Initial Coupon	3.97	3.67
DTI	33.4	33.6
LTV	74.3	74.0
Credit Score	754	753

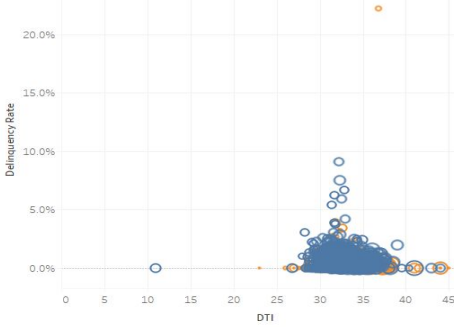
Loan-to-Value vs Initial Loan Amount



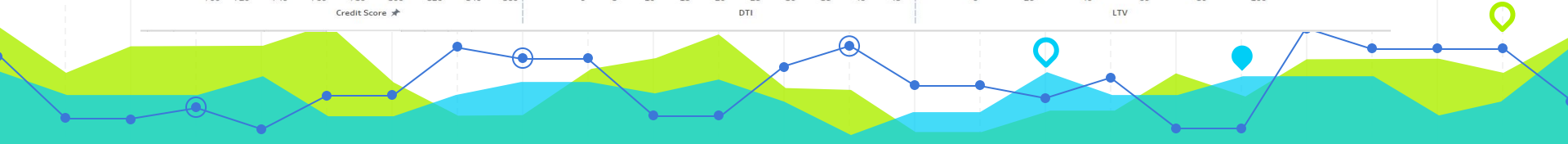
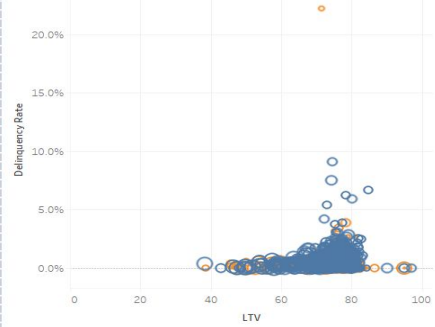
Credit Score vs Delq Rate



DTI vs Delq Rate



Loan-to-Value vs Delq Rate





# Scaling and Extending

## Complete loan performance databases



- 17 years (total) vs 2 years (demo)
- Automated v manual data updates

## Auxiliary housing price, sales and construction data



- FHFA House Price Index (housing price data by zip)
- US Census Bureau (house price index by region)
- Zillow: % homes decreasing / increasing in value (by state)
- Many others, but more problematic.

## Macroeconomic data



- Employment, wages and economic growth
- Demographics



# THANKS!

