February 2020

How Urban Uses Big Data Systems to Democratize Access to LED Data

Vivian Zheng, Data Scientist, Office of Technology and Data Science

URBAN INSTITUTE · ELEVATE · THE · DEBATE

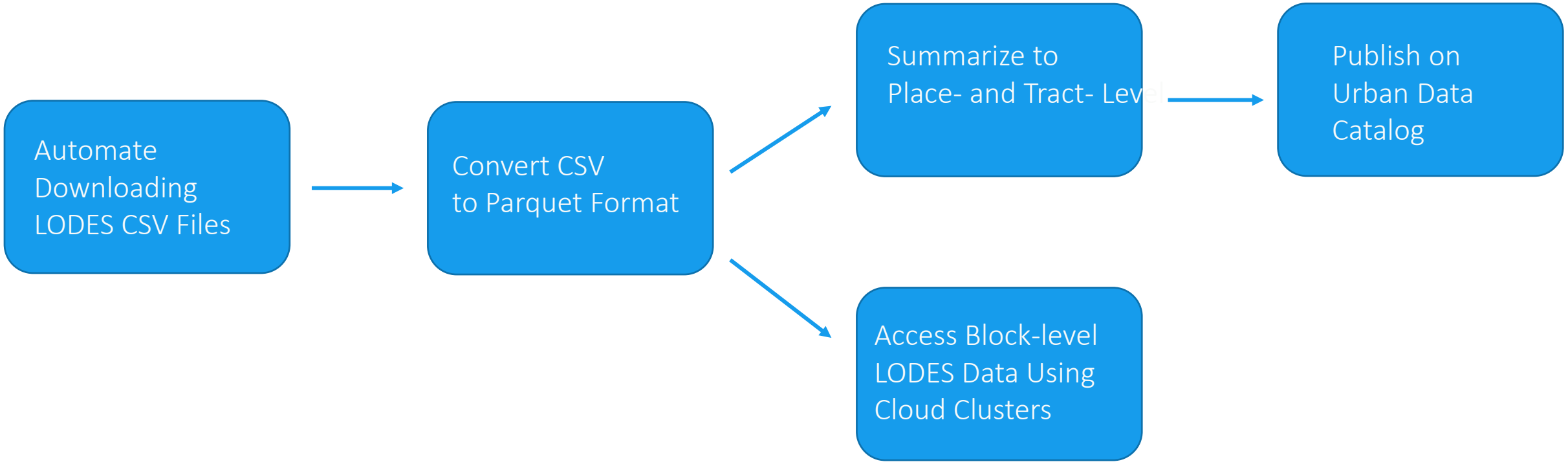
About Urban and the Data Science Team

Easy and Trusted Access to Data

LEHD Origin-Destination Employment Statistics (LODES)

- Block-level LODES Data
  - Detailed spatial distributions of workers' employment and residential locations and the relation between the two
  - In raw form as a set of CSV files
  - Simplify data downloading
- Tract-level and Place-level LODES Data
  - Summarize block-level LODES data to place- and tract-level
  - Available at Urban's Data Catalog (datacatalog.urban.org)

LODES Data Processing Pipeline


Automate Downloading LODES CSV Files
Convert CSV to Parquet Format
Summarize to Place- and Tract- Level
Publish on Urban Data Catalog
Access Block-level LODES Data Using Cloud Clusters

1. Automate downloading LODES files

LEHD Origin-Destination Employment Statistics (LODES)

LEHD Origin-Destination Employment Statistics (LODES) used by OnTheMap are available for download below. Version 7 of LODES was enumerated by 2010 census blocks. Previous versions of LODES were enumerated with 2000 census blocks.

Data files are state-based and organized into three types: Origin-Destination (OD), Residence Area Characteristics (RAC), and Workplace Area Characteristics (WAC), all at census block geographic detail. Data is available for most states for the years 2002–2017.

To browse the LODES data files in their directory structure or to access them with a FTP program (must be able to access HTTP), go to lehd.ces.census.gov/data/lodes/.

Download LODES data*:

Version: LODES7 State/Territory: Virginia

Type: Workplace Area Characteristics (WAC)

View Files Close

Metadata for VA | Geography crosswalk for VA | VA sha256sum file

va_wac_S000_JT00_2002.csv.gz	22 Sep 2017 01:47	959 KB
va_wac_S000_JT00_2003.csv.gz	22 Sep 2017 01:47	952 KB
va_wac_S000_JT00_2004.csv.gz	22 Sep 2017 01:47	970 KB
va_wac_S000_JT00_2005.csv.gz	22 Sep 2017 01:47	1 MB
va_wac_S000_JT00_2006.csv.gz	22 Sep 2017 01:47	1 MB
va_wac_S000_JT00_2007.csv.gz	22 Sep 2017 01:47	1 MB
va_wac_S000_JT00_2008.csv.gz	22 Sep 2017 01:47	1 MB
va_wac_S000_JT00_2009.csv.gz	22 Sep 2017 01:47	2 MB
va_wac_S000_JT00_2010.csv.gz	22 Sep 2017 01:47	2 MB
va_wac_S000_JT00_2011.csv.gz	22 Sep 2017 01:47	1 MB
va_wac_S000_JT00_2012.csv.gz	22 Sep 2017 01:47	1 MB
va_wac_S000_JT00_2013.csv.gz	22 Sep 2017 01:47	1 MB
va_wac_S000_JT00_2014.csv.gz	28 Aug 2019 20:45	1 MB
va_wac_S000_JT00_2015.csv.gz	28 Aug 2019 20:45	1 MB

1. Automate downloading LODES files

https://github.com/UrbanInstitute/lodes-data-downloads

Index of /data/lodes/LODES7

Name	Last modified	Size	Description
Parent Directory		-	
LODESTechDoc7.0.pdf	2017-09-25 09:04	60K	
LODESTechDoc7.1.pdf	2017-09-25 09:04	61K	
LODESTechDoc7.2.pdf	2017-09-25 09:04	60K	
LODESTechDoc7.3.pdf	2017-09-21 21:25	61K	
LODESTechDoc7.4.pdf	2019-08-26 18:16	145K	
ak/	2019-08-28 17:27	-	
al/	2019-08-28 17:27	-	
ar/	2019-08-28 17:27	-	
az/	2019-08-28 17:27	-	
ca/	2019-08-28 17:27	-	
co/	2019-08-28 17:27	-	
ct/	2019-08-28 17:27	-	
dc/	2019-08-28 17:27	-	
de/	2019-08-28 17:27	-	
fl/	2019-08-28 17:27	-	
ga/	2019-08-28 17:27	-	
hi/	2019-08-28 17:27	-	
ia/	2019-08-28 17:27	-	
id/	2019-08-28 17:27	-	
il/	2019-08-28 17:27	-	
in/	2019-08-28 17:27	-	
ks/	2019-08-28 17:27	-	
ky/	2019-08-28 17:27	-	
la/	2019-08-28 17:27	-	
ma/	2019-08-28 17:27	-	
md/	2019-08-28 17:27	-	
me/	2019-08-28 17:27	-	
mi/	2019-08-28 17:27	-	
mn/	2019-08-28 17:27	-	
mo/	2019-08-28 17:27	-	
ms/	2019-08-28 17:27	-	

2. Convert CSV to Parquet Format

a. Compressed File

b. Columnar Storage Format

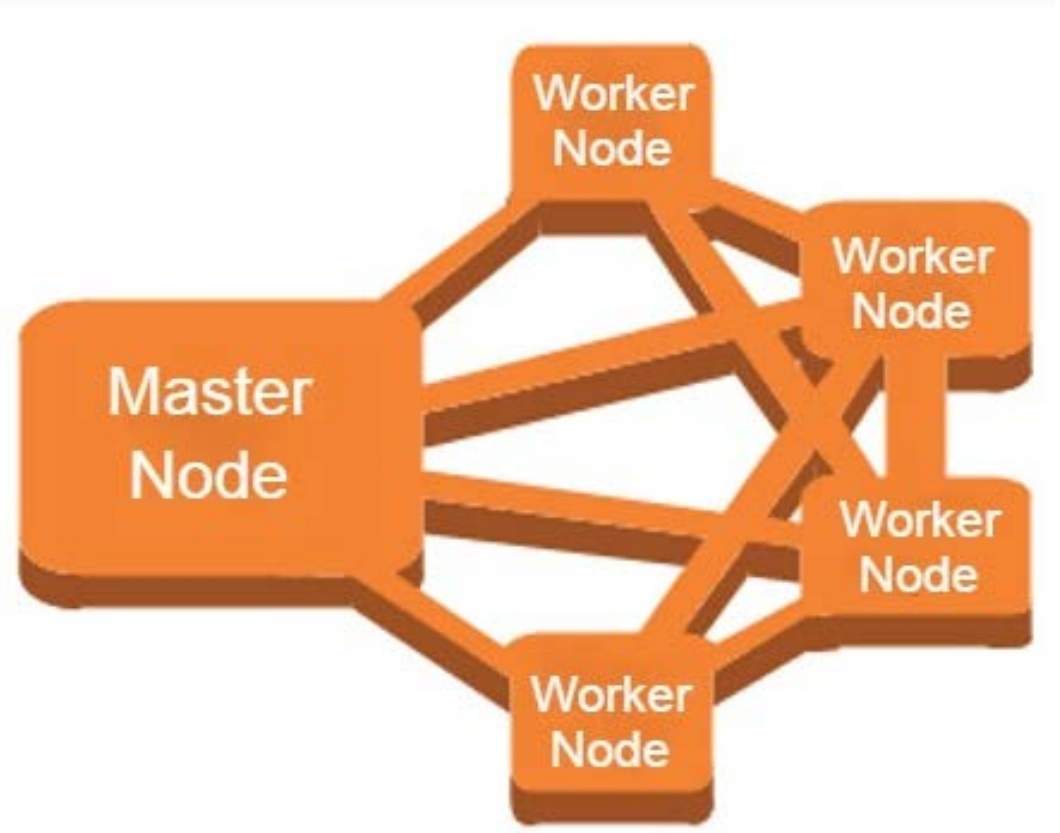
3. Summarize to Place Level and Tract Level

a. Merge with Census geography crosswalks using the Census block number

b. Sum up the total number of workers in each block by Census tract or place number

Cloud Computing Techniques

Cloud Clusters

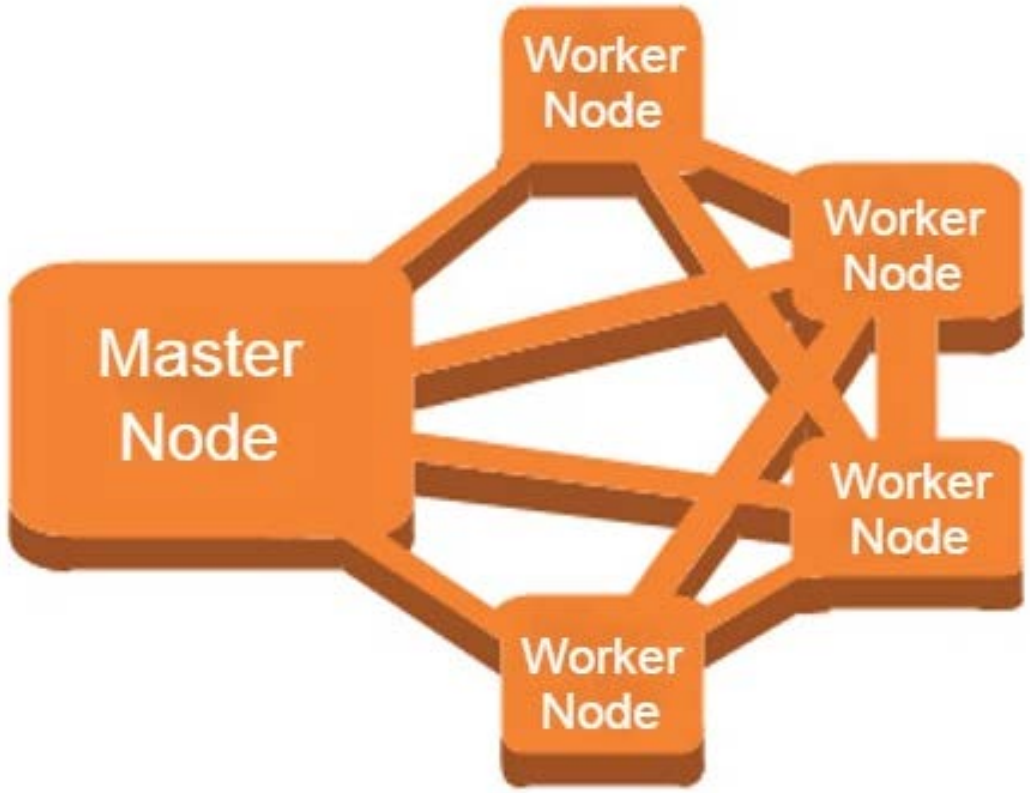


Advantages

- On-demand
- Reduce processing time: 500 hours to 10min
- Cost effective: $2/hour on average
- Scalable
- Spin up more than one cluster at the same time

Cloud Computing Techniques

On-demand Cloud Clusters + Python/R



Datacatalog.urban.org

2 results

Search: lodes | Sort by: Date changed | Order: Descending | Apply | Reset

Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES) Summary Files—Census-Tract Level

Data Technology

Job Market and Labor Force | Neighborhoods, Cities, and Metros

This contains eight LODES census tract level summary files—four Residence Area Characteristics (RAC) and four Workplace Area Characteristics (WAC) files. RAC and WAC data each contain four files representing four job types—all jobs, primary jobs...

8x CSV

Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES) Summary Files—Census-Place Level

Data Technology

Job Market and Labor Force | Neighborhoods, Cities, and Metros

This contains eight LODES census place level summary files—four Residence Area Characteristics (RAC) and four Workplace Area Characteristics (WAC) files. RAC and WAC data each contain four files representing four job types—all jobs, primary jobs...

8x CSV

Longitudinal Employer-Household Dynamics Origin-Destination Employment Statistics (LODES) Summary Files—Census-Tract Level

Job Market and Labor Force | Neighborhoods, Cities, and Metros

This contains eight LODES census tract level summary files—four Residence Area Characteristics (RAC) and four Workplace Area Characteristics (WAC) files. RAC and WAC data each contain four files representing four job types—all jobs, primary jobs, all federal jobs, and federal primary jobs.

Data and Resources

RAC All Jobs Excluding Federal Jobs	Preview	Download
RAC Primary Jobs Excluding Federal Jobs	Preview	Download
WAC All Jobs Excluding Federal Jobs	Preview	Download
WAC Primary Jobs Excluding Federal Jobs	Preview	Download
RAC All Federal Jobs (2010-2015)	Preview	Download
RAC Primary Federal Jobs (2010-2015)	Preview	Download
WAC All Federal Jobs (2010-2015)	Preview	Download
WAC Primary Federal Jobs (2010-2015)	Preview	Download

BIG DATA | BUSINESS | TIME SERIES | SALES | EMPLOYMENT | INDUSTRY | ESTABLISHMENT | CENSUS

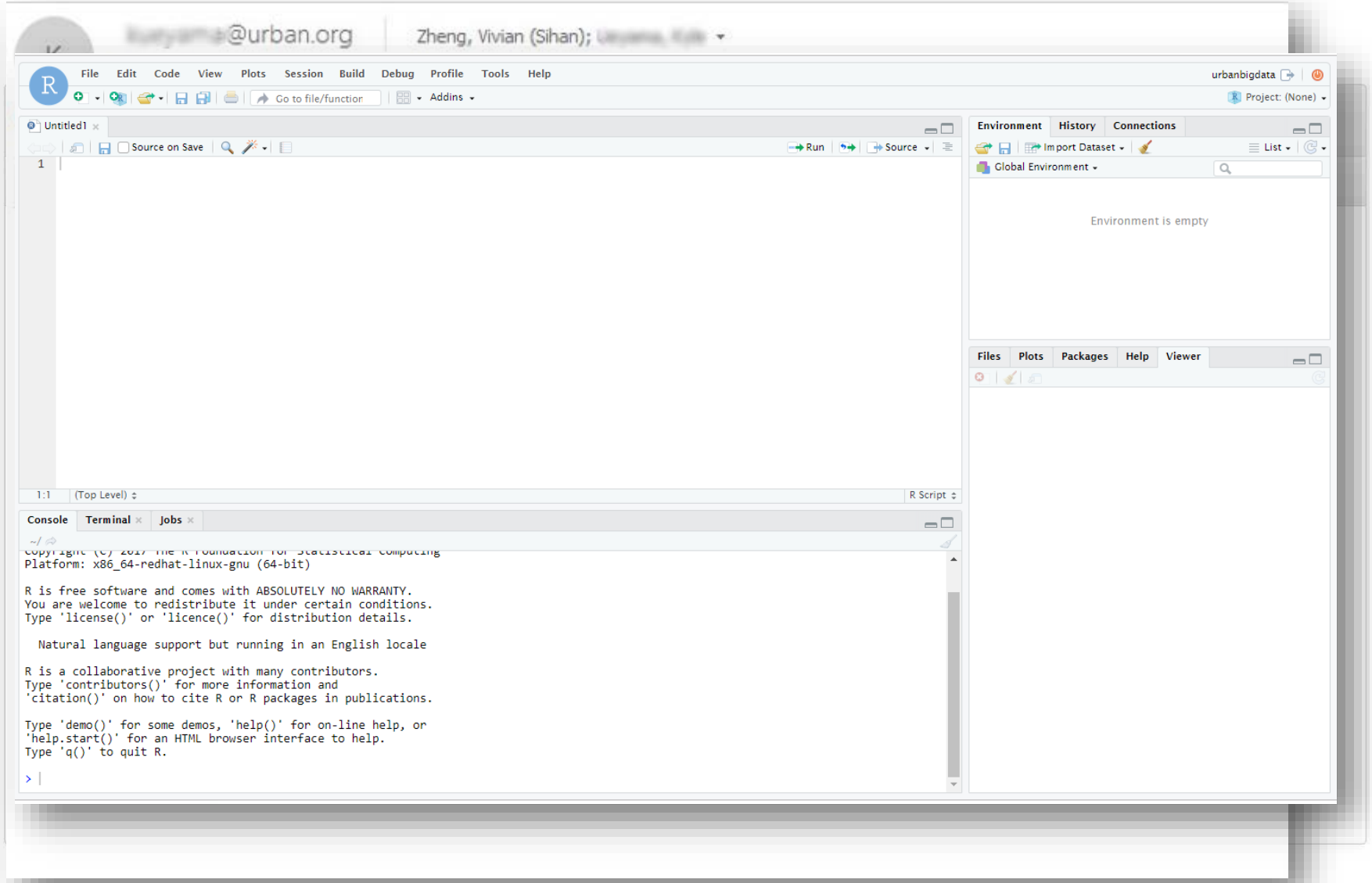
Urban’s Big Data System

Requesting Cloud Clusters

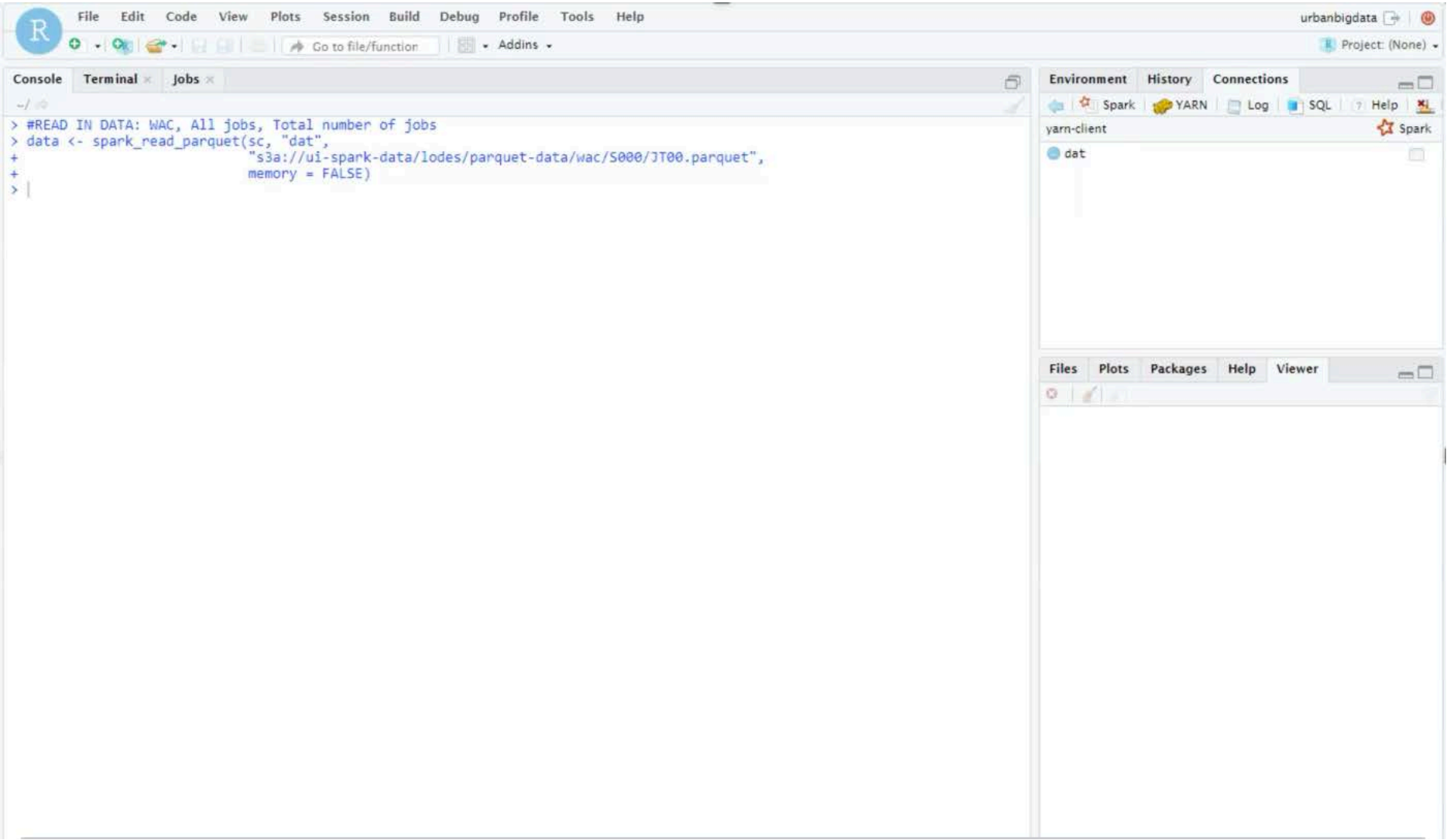
1. Request a cloud cluster

2. Log in through a link from an Email

3. Program in Rstudio-installed web interface



Urban Big Data System - Processing LODES Data in Cloud Cluster



Urban Big Data System - Processing LODES Data in Cloud Cluster

```
> #READ IN DATA: WAC, All jobs, Total number of jobs
> data <- spark_read_parquet(sc, "dat",
+                            "s3a://ui-spark-data/lodes/parquet-data/wac/S000/JT00.parquet",
+                            memory = FALSE)
>
> #show all column names
> colnames(data)
 [1] "w_geocode"  "c000"       "ca01"       "ca02"       "ca03"       "ce01"       "ce02"       "ce03"
 [9] "cns01"      "cns02"      "cns03"      "cns04"      "cns05"      "cns06"      "cns07"      "cns08"
[17] "cns09"      "cns10"      "cns11"      "cns12"      "cns13"      "cns14"      "cns15"      "cns16"
[25] "cns17"      "cns18"      "cns19"      "cns20"      "cr01"       "cr02"       "cr03"       "cr04"
[33] "cr05"       "cr07"       "ct01"       "ct02"       "cd01"       "cd02"       "cd03"       "cd04"
[41] "cs01"       "cs02"       "cfa01"      "cfa02"      "cfa03"      "cfa04"      "cfa05"      "cfs01"
[49] "cfs02"      "cfs03"      "cfs04"      "cfs05"      "createdate" "year"
>
>
> #Count the number of rows
> data_sub <- data %>% select(w_geocode)
> count(data_sub)
# Source: spark<?> [?? x 1]
        n
     <dbl>
1 31829522
>
>
> #View datasets
> dat1 <- data %>%
+     select(total_jobs = c000, w_geocode, year) %>%
+     sdf_sample(fraction = 0.0000001, replacement = FALSE)
>
> dat1
# Source: spark<?> [?? x 3]
  total_jobs w_geocode        year
       <int> <chr>           <int>
1        131 360810171001014  2010
2          2 060590012023013  2013
3         30 391199119001087  2014
4          3 390810010001005  2011
5        162 550250025001083  2010
6         10 410190400003050  2010
>
>
```

```
> dat
  w_geocode : chr "060014001001000" "06"..
  c000 : int 2267 9 16 7474 78
  ca01 : int 367 1 4 1236 10
  ca02 : int 1286 5 7 4275 44
  ca03 : int 614 3 5 1963 24
  ce01 : int 296 0 1 933 11
  ce02 : int 596 3 4 1987 17
  ce03 : int 1375 6 11 4554 50
  cns01 : int 0 0 0 0 0
```

Urban Big Data System - Processing LODES Data in Cloud Cluster

```
+                                memory = FALSE)
>
> #show all column names
> colnames(data)
 [1] "w_geocode"  "c000"       "ca01"       "ca02"       "ca03"       "ce01"       "ce02"       "ce03"
 [9] "cns01"      "cns02"      "cns03"      "cns04"      "cns05"      "cns06"      "cns07"      "cns08"
[17] "cns09"      "cns10"      "cns11"      "cns12"      "cns13"      "cns14"      "cns15"      "cns16"
[25] "cns17"      "cns18"      "cns19"      "cns20"      "cr01"       "cr02"       "cr03"       "cr04"
[33] "cr05"       "cr07"       "ct01"       "ct02"       "cd01"       "cd02"       "cd03"       "cd04"
[41] "cs01"       "cs02"       "cfa01"      "cfa02"      "cfa03"      "cfa04"      "cfa05"      "cfs01"
[49] "cfs02"      "cfs03"      "cfs04"      "cfs05"      "createdate" "year"
>
>
> #Count the number of rows
> data_sub <- data %>% select(w_geocode)
> count(data_sub)
# Source: spark<?> [?? x 1]
         n
     <dbl>
1 31829522
>
>
> #View datasets
> dat1 <- data %>%
+     select(total_jobs = c000, w_geocode, year) %>%
+     sdf_sample(fraction = 0.0000001, replacement = FALSE)
>
> dat1
# Source: spark<?> [?? x 3]
  total_jobs w_geocode        year
       <int> <chr>           <int>
1        131 360810171001014  2010
2          2 060590012023013  2013
3         30 391199119001087  2014
4          3 390810010001005  2011
5        162 550250025001083  2010
6         10 410190400003050  2010
>
>
> #Filter the dataset
> data_2015 <- data %>%
+     filter(year == 2015) %>%
+     select(total_jobs = c000, w_geocode, year)
>
>
> write.csv(data_2015, "lodes_wac_alljobs_2015.csv")
>
```

```
dat
  w_geocode : chr "060014001001000" "06"..
  c000 : int 2267 9 16 7474 78
  ca01 : int 367 1 4 1236 10
  ca02 : int 1286 5 7 4275 44
  ca03 : int 614 3 5 1963 24
  ce01 : int 296 0 1 933 11
  ce02 : int 596 3 4 1987 17
  ce03 : int 1375 6 11 4554 50
  cns01 : int 0 0 0 0 0
```

Export Files

The selected file(s) will be downloaded to your computer. Please specify a name for the downloaded file:

lodes_wac_alljobs_2015.csv

Download Cancel

Name	Size	Modif
03_subsetting.md	12.5 KB	Aug
04_missing-data.md	16.9 KB	Aug
05_summary-statistics.md	15 KB	Aug
06_merging.md	18.4 KB	Aug
07_visualizations.md	14 KB	Aug
08_databases-with-jdbc.md	15.7 KB	Aug
09_glm.md	25.6 KB	Aug
10_timeseries-1.md	21.2 KB	Aug
derby.log	681 B	Aug
glm_files		
License.md	34.4 KB	Aug
lodes_wac_alljobs_2015.csv	71.4 MB	Aug
metastore_db		
R		
README.md	16.1 KB	Aug
rmd		
spark		
visualizations_files		

Future work

- Update new LODES data
- Include more job types for RAC and WAC
- Include LODES Origin-Destination Data
- Integrate more big datasets in Urban Data Catalog

  (Datacatalog.urban.org)

THANK YOU!