Implementation of a Data Engineering ETL Pipeline using Databricks

Ingestion, Transformation, Aggregation, and Gold-Level Analytics

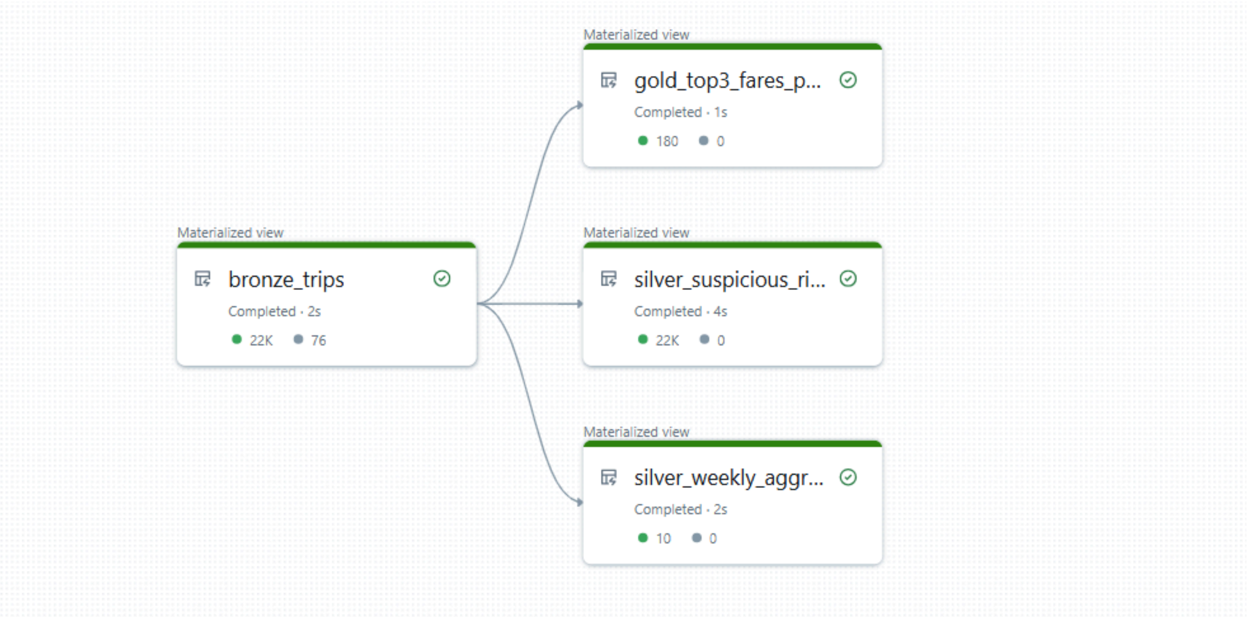
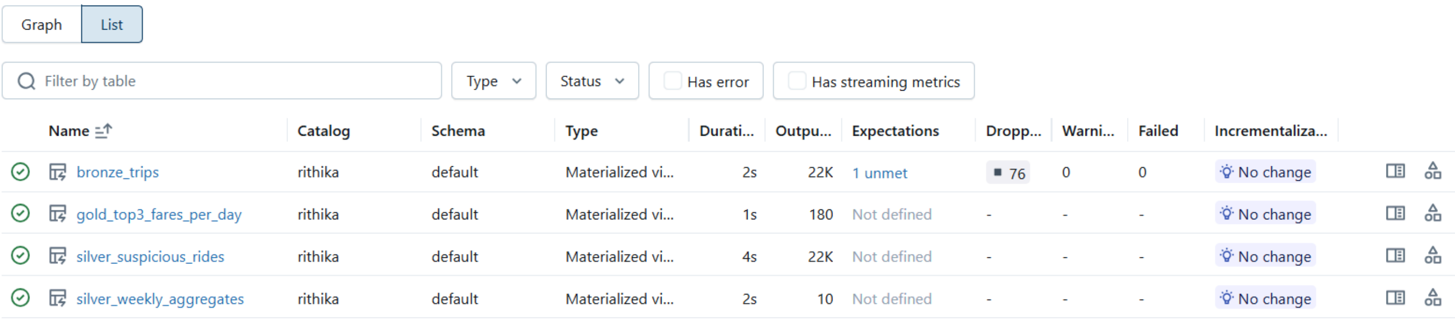
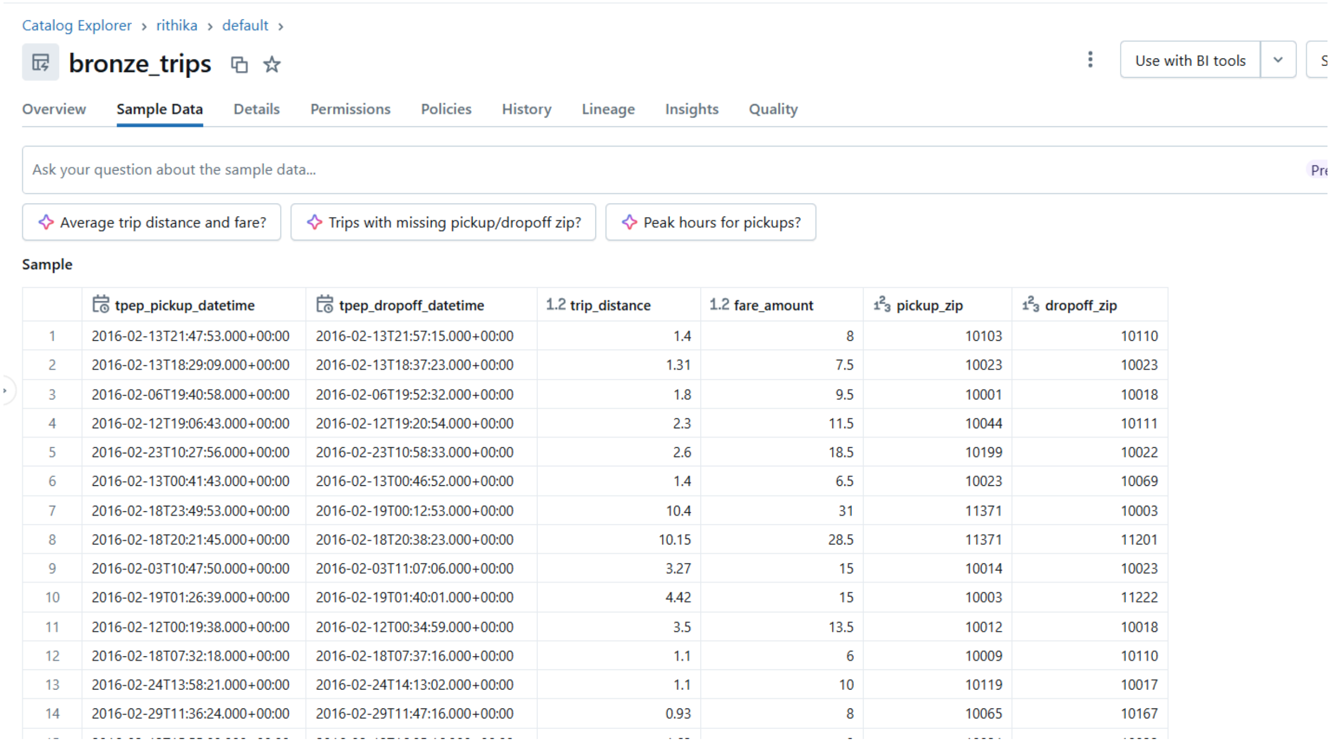


Figure 1: Lakehouse pipeline graph showing data flow from the Bronze layer (raw trips) to Silver tables (suspicious rides & weekly aggregates) and finally to the Gold materialized view of top-3 highest fare rides per day.



Name	Catalog	Schema	Type	Durati...	Outpu...	Expectations	Dropp...	Warni...	Failed	Incrementaliza...
bronze_trips	rithika	default	Materialized vi...	2s	22K	1 unmet	■ 76	0	0	No change
gold_top3_fares_per_day	rithika	default	Materialized vi...	1s	180	Not defined	-	-	-	No change
silver_suspicious_rides	rithika	default	Materialized vi...	4s	22K	Not defined	-	-	-	No change
silver_weekly_aggregates	rithika	default	Materialized vi...	2s	10	Not defined	-	-	-	No change

Figure 2: List view of all Delta Live Tables showing table type, status, output row counts, expectations, and incremental processing details.



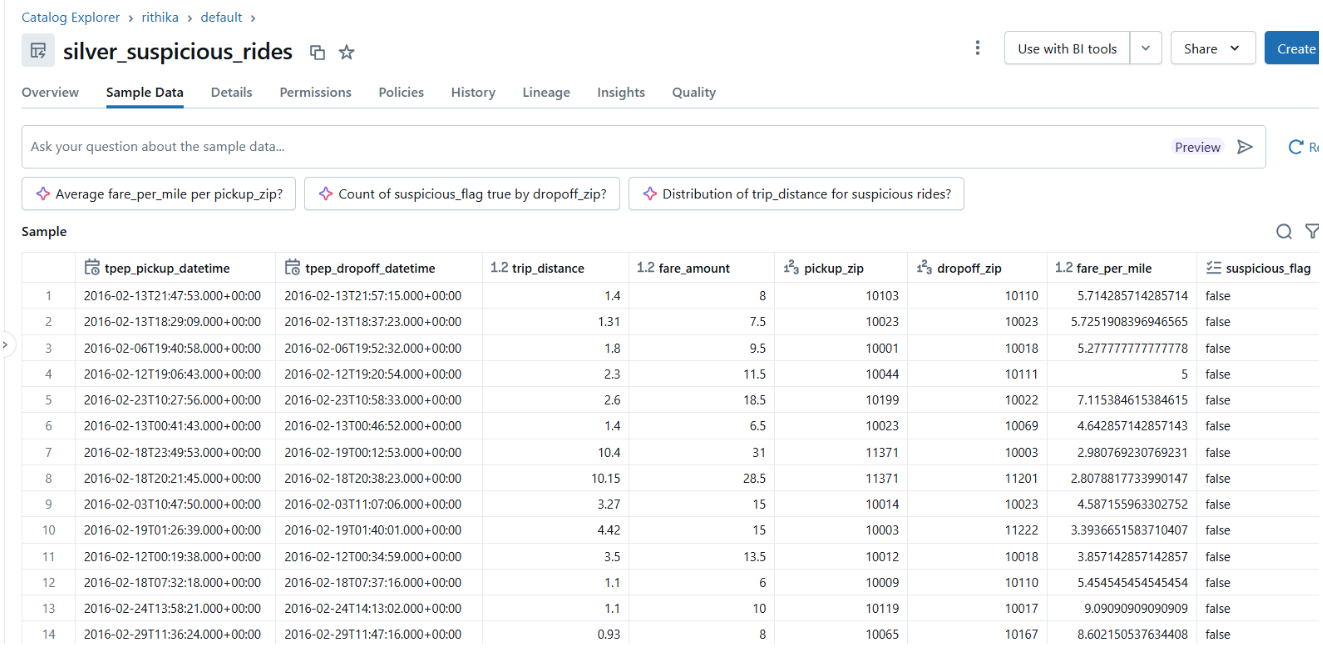
Catalog Explorer › rithika › default ›

bronze_trips

Sample

	tpep_pickup_datetime	tpep_dropoff_datetime	trip_distance	fare_amount	pickup_zip	dropoff_zip
1	2016-02-13T21:47:53.000+00:00	2016-02-13T21:57:15.000+00:00	1.4	8	10103	10110
2	2016-02-13T18:29:09.000+00:00	2016-02-13T18:37:23.000+00:00	1.31	7.5	10023	10023
3	2016-02-06T19:40:58.000+00:00	2016-02-06T19:52:32.000+00:00	1.8	9.5	10001	10018
4	2016-02-12T19:06:43.000+00:00	2016-02-12T19:20:54.000+00:00	2.3	11.5	10044	10111
5	2016-02-23T10:27:56.000+00:00	2016-02-23T10:58:33.000+00:00	2.6	18.5	10199	10022
6	2016-02-13T00:41:43.000+00:00	2016-02-13T00:46:52.000+00:00	1.4	6.5	10023	10069
7	2016-02-18T23:49:53.000+00:00	2016-02-19T00:12:53.000+00:00	10.4	31	11371	10003
8	2016-02-18T20:21:45.000+00:00	2016-02-18T20:38:23.000+00:00	10.15	28.5	11371	11201
9	2016-02-03T10:47:50.000+00:00	2016-02-03T11:07:06.000+00:00	3.27	15	10014	10023
10	2016-02-19T01:26:39.000+00:00	2016-02-19T01:40:01.000+00:00	4.42	15	10003	11222
11	2016-02-12T00:19:38.000+00:00	2016-02-12T00:34:59.000+00:00	3.5	13.5	10012	10018
12	2016-02-18T07:32:18.000+00:00	2016-02-18T07:37:16.000+00:00	1.1	6	10009	10110
13	2016-02-24T13:58:21.000+00:00	2016-02-24T14:13:02.000+00:00	1.1	10	10119	10017
14	2016-02-29T11:36:24.000+00:00	2016-02-29T11:47:16.000+00:00	0.93	8	10065	10167

Figure 3: Sample data from the Bronze layer (bronze_trips) containing cleaned and type-cast raw taxi trip records.



Catalog Explorer › rithika › default ›

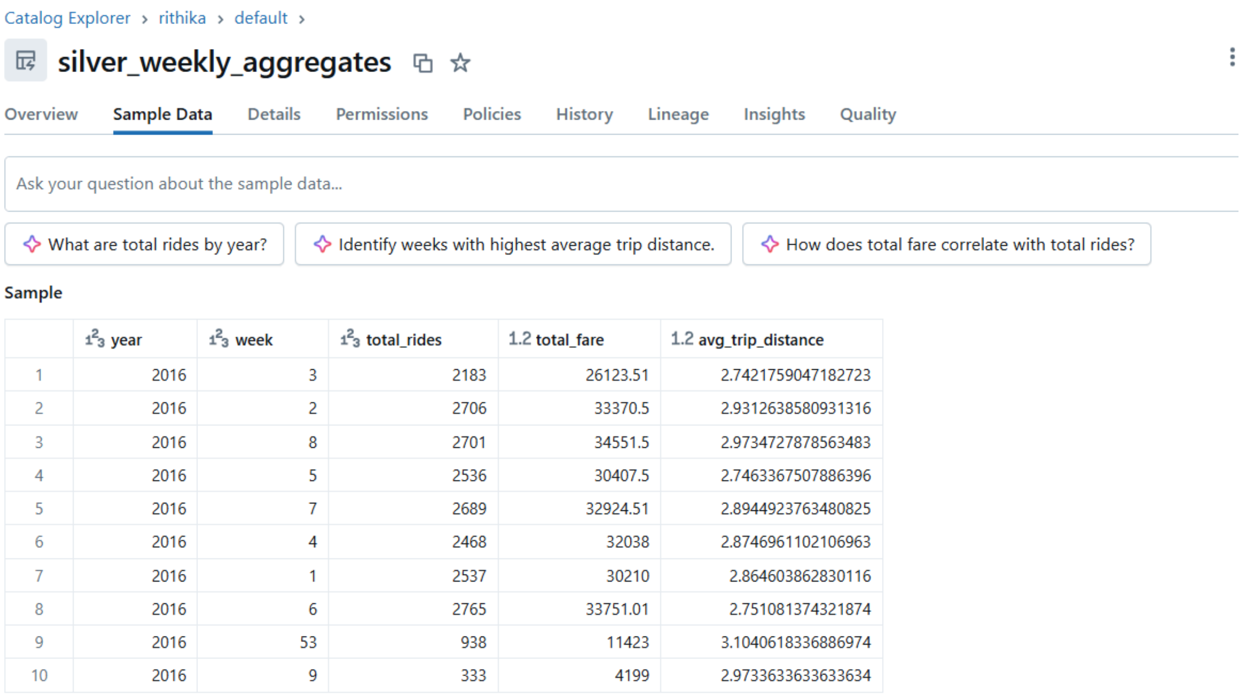
silver_suspicious_rides

Use with BI tools
Share
Create
Overview
Sample Data
Details
Permissions
Policies
History
Lineage
Insights
Quality
Ask your question about the sample data...
Preview
Average fare_per_mile per pickup_zip?
Count of suspicious_flag true by dropoff_zip?
Distribution of trip_distance for suspicious rides?

Sample

	tpep_pickup_datetime	tpep_dropoff_datetime	trip_distance	fare_amount	pickup_zip	dropoff_zip	fare_per_mile	suspicious_flag
1	2016-02-13T21:47:53.000+00:00	2016-02-13T21:57:15.000+00:00	1.4	8	10103	10110	5.714285714285714	false
2	2016-02-13T18:29:09.000+00:00	2016-02-13T18:37:23.000+00:00	1.31	7.5	10023	10023	5.7251908396946565	false
3	2016-02-06T19:40:58.000+00:00	2016-02-06T19:52:32.000+00:00	1.8	9.5	10001	10018	5.277777777777778	false
4	2016-02-12T19:06:43.000+00:00	2016-02-12T19:20:54.000+00:00	2.3	11.5	10044	10111	5	false
5	2016-02-23T10:27:56.000+00:00	2016-02-23T10:58:33.000+00:00	2.6	18.5	10199	10022	7.115384615384615	false
6	2016-02-13T00:41:43.000+00:00	2016-02-13T00:46:52.000+00:00	1.4	6.5	10023	10069	4.642857142857143	false
7	2016-02-18T23:49:53.000+00:00	2016-02-19T00:12:53.000+00:00	10.4	31	11371	10003	2.980769230769231	false
8	2016-02-18T20:21:45.000+00:00	2016-02-18T20:38:23.000+00:00	10.15	28.5	11371	11201	2.8078817733990147	false
9	2016-02-03T10:47:50.000+00:00	2016-02-03T11:07:06.000+00:00	3.27	15	10014	10023	4.587155963302752	false
10	2016-02-19T01:26:39.000+00:00	2016-02-19T01:40:01.000+00:00	4.42	15	10003	11222	3.3936651583710407	false
11	2016-02-12T00:19:38.000+00:00	2016-02-12T00:34:59.000+00:00	3.5	13.5	10012	10018	3.857142857142857	false
12	2016-02-18T07:32:18.000+00:00	2016-02-18T07:37:16.000+00:00	1.1	6	10009	10110	5.454545454545454	false
13	2016-02-24T13:58:21.000+00:00	2016-02-24T14:13:02.000+00:00	1.1	10	10119	10017	9.09090909090909	false
14	2016-02-29T11:36:24.000+00:00	2016-02-29T11:47:16.000+00:00	0.93	8	10065	10167	8.602150537634408	false

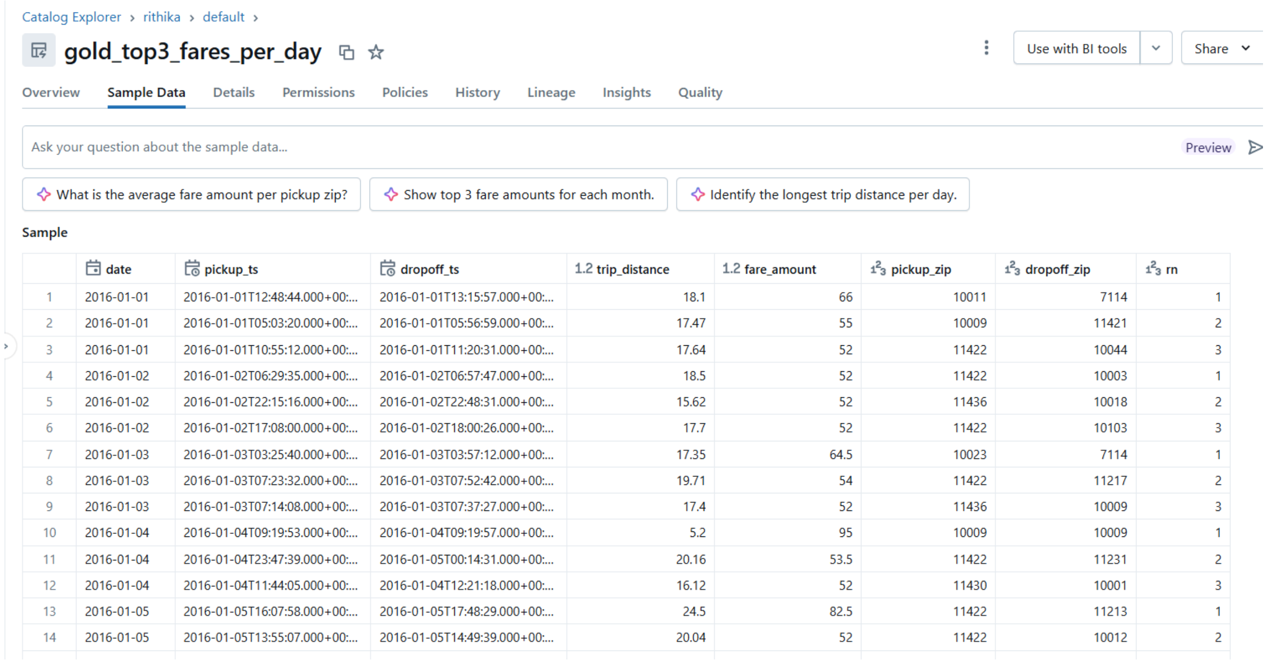
Figure 4: Sample data from the Silver table (silver_suspicious_rides) showing calculated fare_per_mile and the suspicious ride flag.


Catalog Explorer › rithika › default ›
silver_weekly_aggregates
Overview
Sample Data
Details
Permissions
Policies
History
Lineage
Insights
Quality
Ask your question about the sample data...
What are total rides by year?
Identify weeks with highest average trip distance.
How does total fare correlate with total rides?

Sample

	year	week	total_rides	total_fare	avg_trip_distance
1	2016	3	2183	26123.51	2.7421759047182723
2	2016	2	2706	33370.5	2.9312638580931316
3	2016	8	2701	34551.5	2.9734727878563483
4	2016	5	2536	30407.5	2.7463367507886396
5	2016	7	2689	32924.51	2.8944923763480825
6	2016	4	2468	32038	2.8746961102106963
7	2016	1	2537	30210	2.864603862830116
8	2016	6	2765	33751.01	2.751081374321874
9	2016	53	938	11423	3.1040618336886974
10	2016	9	333	4199	2.9733633633633634

Figure 5: Sample data from the Silver aggregated table (silver_weekly_aggregates) summarizing weekly ride counts, total fare, and average trip distance.


Catalog Explorer › rithika › default ›
gold_top3_fares_per_day
Use with BI tools
Share
Overview
Sample Data
Details
Permissions
Policies
History
Lineage
Insights
Quality
Ask your question about the sample data...
Preview
What is the average fare amount per pickup zip?
Show top 3 fare amounts for each month.
Identify the longest trip distance per day.

Sample

	date	pickup_ts	dropoff_ts	trip_distance	fare_amount	pickup_zip	dropoff_zip	rn
1	2016-01-01	2016-01-01T12:48:44.000+00:...	2016-01-01T13:15:57.000+00:...	18.1	66	10011	7114	1
2	2016-01-01	2016-01-01T05:03:20.000+00:...	2016-01-01T05:56:59.000+00:...	17.47	55	10009	11421	2
3	2016-01-01	2016-01-01T10:55:12.000+00:...	2016-01-01T11:20:31.000+00:...	17.64	52	11422	10044	3
4	2016-01-02	2016-01-02T06:29:35.000+00:...	2016-01-02T06:57:47.000+00:...	18.5	52	11422	10003	1
5	2016-01-02	2016-01-02T22:15:16.000+00:...	2016-01-02T22:48:31.000+00:...	15.62	52	11436	10018	2
6	2016-01-02	2016-01-02T17:08:00.000+00:...	2016-01-02T18:00:26.000+00:...	17.7	52	11422	10103	3
7	2016-01-03	2016-01-03T03:25:40.000+00:...	2016-01-03T03:57:12.000+00:...	17.35	64.5	10023	7114	1
8	2016-01-03	2016-01-03T07:23:32.000+00:...	2016-01-03T07:52:42.000+00:...	19.71	54	11422	11217	2
9	2016-01-03	2016-01-03T07:14:08.000+00:...	2016-01-03T07:37:27.000+00:...	17.4	52	11436	10009	3
10	2016-01-04	2016-01-04T09:19:53.000+00:...	2016-01-04T09:19:57.000+00:...	5.2	95	10009	10009	1
11	2016-01-04	2016-01-04T23:47:39.000+00:...	2016-01-05T00:14:31.000+00:...	20.16	53.5	11422	11231	2
12	2016-01-04	2016-01-04T11:44:05.000+00:...	2016-01-04T12:21:18.000+00:...	16.12	52	11430	10001	3
13	2016-01-05	2016-01-05T16:07:58.000+00:...	2016-01-05T17:48:29.000+00:...	24.5	82.5	11422	11213	1
14	2016-01-05	2016-01-05T13:55:07.000+00:...	2016-01-05T14:49:39.000+00:...	20.04	52	11422	10012	2

Figure 6: Sample data from the Gold materialized view (gold_top3_fares_per_day) showing the top-3 highest fare rides for each date.