Operation Analytics - Case Study 1

Project Description

Operational Analytics is a critical process aimed at improving end-to-end operations within a company. As the Lead Data Analyst in this project, the objective was to analyze a dataset named `job_data` to derive insights that could aid different departments in understanding and optimizing their operations. The primary focus areas included investigating metric spikes, calculating throughput, and analyzing language distribution.

Approach

The analysis was conducted using advanced SQL skills, leveraging the dataset's key components, such as job reviews, throughput metrics, and language information. The approach involved writing SQL queries to answer specific questions posed by different departments. The tasks included calculating jobs reviewed over time, determining the 7-day rolling average of throughput, and analyzing the percentage share of each language in the last 30 days. Additionally, a query was crafted to identify duplicate rows in the `job_data` table.

To enhance the dataset's richness and complexity, DDL commands were employed to add supplementary data to the `job_data` table. This demonstrated proficiency in Data Definition Language. The added data contributed to a more comprehensive analysis, allowing for a deeper exploration of trends and patterns within the operational metrics.

```
# Changing datatype of 'ds' column to DATE
UPDATE
    job_data
SET
    ds = str_to_date(ds, '%m/%d/%Y');

ALTER TABLE
    job_data modify COLUMN ds DATE;
```

```
#Inserting new data
INSERT INTO
    job_data (ds, job_id, actor_id, event, LANGUAGE, time_spent, org)
VALUES
    ('2020-12-01', 27, 1012, 'skip', 'English', 67, 'C'),
    ('2020-12-01', 29, 1039, 'decision', 'French', 58, 'A'),
    ('2020-12-02', 23, 1013, 'transfer', 'Arabic', 20, 'D'),
    ('2020-12-03', 26, 1032, 'skip', 'Persian', 70, 'B'),
    ('2020-12-03', 30, 1043, 'decision', 'Hindi', 85, 'A'),
    ('2020-12-04', 28, 1044, 'transfer', 'Italian', 36, 'C'),
    ('2020-12-05', 25, 1033, 'skip', 'English', 15, 'D'),
    ('2020-12-06', 29, 1011, 'decision', 'French', 92, 'B'),
    ('2020-12-07', 31, 1035, 'transfer', 'Arabic', 54, 'A'),
    ('2020-12-08', 28, 1047, 'skip', 'Persian', 32, 'C'),
    ('2020-12-08', 32, 1037, 'decision', 'Hindi', 97, 'D'),
    ('2020-12-09', 30, 1036, 'transfer', 'Italian', 29, 'B'),
    ('2020-12-10', 26, 1014, 'skip', 'English', 63, 'A'),
    ('2020-12-10', 33, 1057, 'decision', 'French', 26, 'C'),
    ('2020-12-11', 29, 1045, 'transfer', 'Arabic', 11, 'D'),
    ('2020-12-12', 28, 1025, 'skip', 'Persian', 70, 'B'),
    ('2020-12-13', 31, 1015, 'decision', 'Hindi', 15, 'A'),
    ('2020-12-13', 34, 1059, 'transfer', 'Italian', 62, 'C'),
    ('2020-12-14', 30, 1035, 'skip', 'English', 52, 'D'),
    ('2020-12-15', 27, 1013, 'decision', 'French', 35, 'B'),
    ('2020-12-15', 40, 1043, 'transfer', 'Arabic', 53, 'A'),
    ('2020-12-16', 29, 1041, 'skip', 'Persian', 96, 'C'),
    ('2020-12-17', 28, 1061, 'decision', 'Hindi', 83, 'D'),
```

```
    ('2020-12-18', 32, 1031, 'transfer', 'Italian', 44, 'B'),
    ('2020-12-18', 35, 1053, 'skip', 'English', 77, 'A'),
    ('2020-12-19', 31, 1015, 'decision', 'French', 59, 'C'),
    ('2020-12-20', 27, 1027, 'transfer', 'Arabic', 66, 'D'),
    ('2020-12-21', 30, 1063, 'skip', 'Persian', 14, 'B'),
    ('2020-12-21', 33, 1039, 'decision', 'Hindi', 21, 'A'),
    ('2020-12-22', 29, 1065, 'transfer', 'Italian', 38, 'C'),
    ('2020-12-23', 31, 1029, 'skip', 'English', 95, 'D'),
    ('2020-12-24', 28, 1011, 'decision', 'French', 23, 'B'),
    ('2020-12-24', 42, 1045, 'transfer', 'Arabic', 42, 'A'),
    ('2020-12-25', 32, 1073, 'skip', 'Persian', 71, 'C'),
    ('2020-12-26', 31, 1049, 'decision', 'Hindi', 89, 'D'),
    ('2020-12-26', 44, 1047, 'transfer', 'Italian', 100, 'B'),
    ('2020-12-27', 34, 1013, 'skip', 'English', 55, 'A'),
    ('2020-12-28', 32, 1025, 'decision', 'French', 17, 'C'),
    ('2020-12-29', 35, 1017, 'transfer', 'Arabic', 84, 'D'),
    ('2020-12-29', 38, 1033, 'skip', 'Persian', 40, 'B'),
    ('2020-12-30', 32, 1061, 'decision', 'Hindi', 27, 'A'),
    ('2020-12-31', 36, 1019, 'transfer', 'Italian', 78, 'C'),
    ('2020-12-31', 33, 1047, 'skip', 'English', 75, 'D'),
    ('2021-01-01', 37, 1021, 'decision', 'French', 33, 'B'),
    ('2021-01-01', 40, 1031, 'transfer', 'Arabic', 60, 'A'),
    ('2021-01-02', 34, 1067, 'skip', 'Persian', 48, 'C'),
    ('2021-01-03', 33, 1087, 'decision', 'Hindi', 105, 'D'),
    ('2021-01-03', 46, 1045, 'transfer', 'Italian', 80, 'B'),
    ('2021-01-04', 35, 1053, 'skip', 'English', 41, 'A'),
    ('2021-01-05', 34, 1037, 'decision', 'French', 65, 'C'),
    ('2021-01-06', 39, 1089, 'transfer', 'Arabic', 12, 'D'),
    ('2021-01-06', 42, 1031, 'skip', 'Persian', 69, 'B'),
    ('2021-01-07', 36, 1013, 'decision', 'Hindi', 83, 'A'),
    ('2021-01-08', 35, 1045, 'transfer', 'Italian', 24, 'C'),
    ('2021-01-08', 44, 1095, 'skip', 'English', 87, 'D'),
    ('2021-01-09', 36, 1027, 'decision', 'French', 51, 'B'),
    ('2021-01-10', 41, 1097, 'transfer', 'Arabic', 68, 'A'),
    ('2021-01-10', 48, 1061, 'skip', 'Persian', 76, 'C'),
    ('2021-01-11', 38, 1101, 'decision', 'Hindi', 39, 'D'),
    ('2021-01-12', 42, 1039, 'transfer', 'Italian', 56, 'B'),
    ('2021-01-12', 51, 1049, 'skip', 'English', 103, 'A'),
    ('2021-01-13', 38, 1031, 'decision', 'French', 61, 'C'),
    ('2021-01-14', 43, 1103, 'transfer', 'Arabic', 10, 'D'),
    ('2021-01-14', 54, 1083, 'skip', 'Persian', 91, 'B'),
    ('2021-01-15', 39, 1035, 'decision', 'Hindi', 29, 'A'),
    ('2021-01-16', 44, 1047, 'transfer', 'Italian', 88, 'C'),
    ('2021-01-16', 57, 1105, 'skip', 'English', 45, 'D'),
    ('2021-01-17', 41, 1013, 'decision', 'French', 93, 'B'),
    ('2021-01-18', 45, 1107, 'transfer', 'Arabic', 72, 'A'),
    ('2021-01-18', 60, 1067, 'skip', 'Persian', 79, 'C'),
    ('2021-01-19', 42, 1111, 'decision', 'Hindi', 57, 'D'),
    ('2021-01-20', 46, 1045, 'transfer', 'Italian', 64, 'B'),
    ('2021-01-20', 63, 1113, 'skip', 'English', 12, 'A'),
    ('2021-01-21', 43, 1015, 'decision', 'French', 81, 'C'),
    ('2021-01-22', 47, 1117, 'transfer', 'Arabic', 50, 'D'),
    ('2021-01-22', 66, 1087, 'skip', 'Persian', 17, 'B'),
    ('2021-01-23', 44, 1023, 'decision', 'Hindi', 101, 'A'),
    ('2021-01-24', 40, 1033, 'transfer', 'Italian', 32, 'C'),
    ('2021-01-24', 69, 1119, 'skip', 'English', 94, 'D'),
    ('2021-01-25', 43, 1017, 'decision', 'French', 104, 'B'),
    ('2021-01-26', 47, 1121, 'transfer', 'Arabic', 37, 'A'),
```

```
  ('2021-01-26', 72, 1111, 'decision', 'Persian', 21, 'C'),
  ('2021-01-27', 46, 1037, 'transfer', 'Italian', 77, 'D'),
  ('2021-01-28', 43, 1053, 'skip', 'English', 102, 'B'),
  ('2021-01-28', 75, 1123, 'decision', 'French', 90, 'A'),
  ('2021-01-29', 47, 1125, 'transfer', 'Arabic', 49, 'C'),
  ('2021-01-30', 46, 1067, 'skip', 'Persian', 45, 'D'),
  ('2021-01-30', 78, 1127, 'decision', 'Hindi', 74, 'B'),
  ('2021-01-31', 48, 1053, 'transfer', 'Italian', 33, 'A'),
  ('2021-02-01', 45, 1129, 'skip', 'English', 70, 'C'),
  ('2021-02-01', 81, 1131, 'decision', 'French', 99, 'D'),
  ('2021-02-02', 49, 1135, 'transfer', 'Arabic', 106, 'B'),
  ('2021-02-03', 48, 1087, 'skip', 'Persian', 82, 'A'),
  ('2021-02-03', 84, 1137, 'decision', 'Hindi', 52, 'C'),
  ('2021-02-04', 51, 1139, 'transfer', 'Italian', 19, 'D'),
  ('2021-02-05', 49, 1141, 'skip', 'English', 98, 'B'),
  ('2021-02-06', 48, 1023, 'decision', 'French', 107, 'A'),
  ('2021-02-06', 87, 1143, 'transfer', 'Arabic', 25, 'C'),
  ('2021-02-07', 52, 1145, 'skip', 'Persian', 42, 'D'),
  ('2021-02-08', 51, 1013, 'decision', 'Hindi', 84, 'B'),
  ('2021-02-08', 90, 1147, 'transfer', 'Italian', 30, 'A'),
  ('2021-02-09', 52, 1149, 'skip', 'English', 76, 'C'),
  ('2021-02-10', 51, 1017, 'decision', 'French', 97, 'D'),
  ('2021-02-10', 93, 1151, 'transfer', 'Arabic', 61, 'B'),
  ('2021-02-11', 55, 1153, 'skip', 'Persian', 67, 'A'),
  ('2021-02-12', 54, 1037, 'decision', 'Hindi', 92, 'C'),
  ('2021-02-12', 96, 1155, 'transfer', 'Italian', 73, 'D'),
  ('2021-02-13', 56, 1157, 'skip', 'English', 22, 'B'),
  ('2021-02-14', 55, 1021, 'decision', 'French', 59, 'A'),
  ('2021-02-15', 40, 1033, 'transfer', 'Arabic', 32, 'C');
```

Tech-Stack Used

The following software and versions were utilized for the successful execution of the project:

- **MySQL Workbench, version 8.0 CE:**

  MySQL Workbench played a pivotal role in the initial stages of the project. It was primarily employed for creating the database structure. The features of MySQL Workbench ensured a robust foundation for the subsequent data analysis.

- **Beekeeper Studio, version 4.0.3 CE:**

  For query editing and advanced SQL analysis, Beekeeper Studio was employed. Its user-friendly interface and quality-of-life improvements over MySQL Workbench provided a seamless experience during the query development process. The simplistic UI enhanced productivity, allowing for efficient exploration of the dataset and execution of complex SQL queries.

Apart from database and query editing software, MS Excel was employed to produce plots from the query results.

Insights

Jobs Reviewed Over Time:

```
SELECT
    ds AS DATE,
    SUM(time_spent) / 3600 AS hours,
    COUNT(job_id) AS num_jobs,
    ROUND(COUNT(job_id) / SUM(time_spent) * 3600, 2) AS jobs_per_hour
FROM
    job_data
WHERE
    ds BETWEEN '2020-11-01' AND '2020-11-30'
GROUP BY
```

```
    ds
ORDER BY
    DATE;
```

DATE	hours	num_jobs	jobs_per_hour
2020-11-25	0.0125	1	80.00
2020-11-26	0.0156	1	64.29
2020-11-27	0.0289	1	34.62
2020-11-28	0.0092	2	218.18
2020-11-29	0.0056	1	180.00
2020-11-30	0.0111	2	180.00

- The analysis revealed variations in the number of jobs reviewed per hour, providing insights into the pace of job reviews on different days in November 2020.
- This information can be valuable for identifying trends and optimizing the allocation of resources for job reviews.

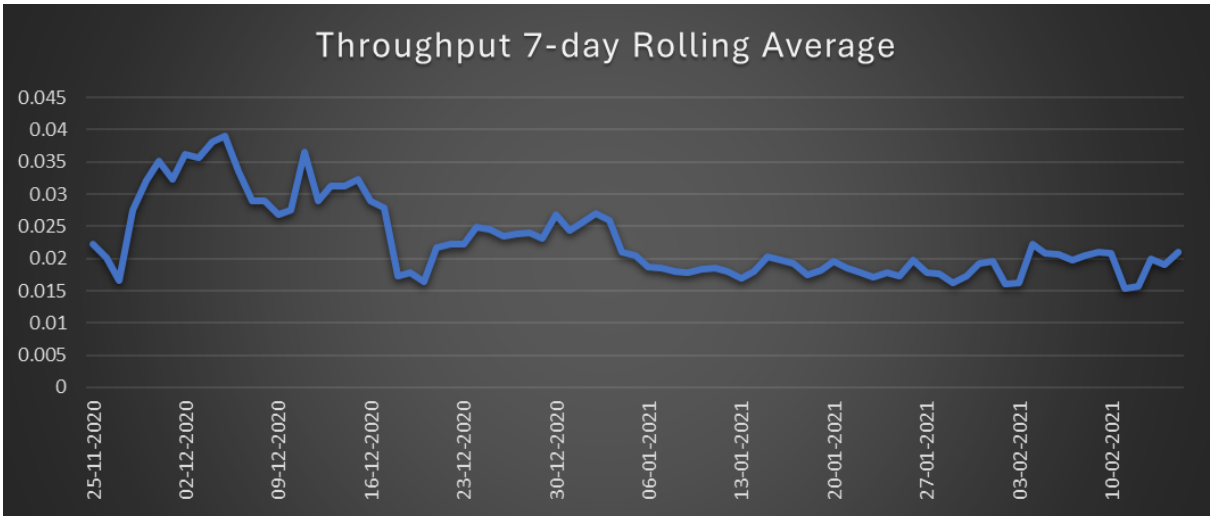
Throughput Analysis:

```
WITH daily_events AS (
    SELECT
        ds,
        COUNT(event) / SUM(time_spent) AS throughput
    FROM
        job_data
    GROUP BY
        ds
)
SELECT
    ds AS DATE,
    AVG(throughput) OVER (
        ORDER BY
            ds ROWS BETWEEN 6 PRECEDING
            AND CURRENT ROW
    ) AS throughput_rolling_avg # 7-day rolling avg
FROM
    daily_events;
```

DATE	throughput_rolling_avg
2020-11-25	0.02220000
2020-11-26	0.02005000
2020-11-27	0.01656667
2020-11-28	0.02757500
2020-11-29	0.03206000
2020-11-30	0.03505000
2020-12-01	0.03224286
2020-12-02	0.03621429
2020-12-03	0.03560000
2020-12-04	0.03820000
2020-12-05	0.03907143
2020-12-06	0.03348571
2020-12-07	0.02898571
2020-12-08	0.02900000

DATE	throughput_rolling_avg
2020-12-09	0.02678571
2020-12-10	0.02750000
2020-12-11	0.03651429
2020-12-12	0.02902857
2020-12-13	0.03118571
2020-12-14	0.03128571
2020-12-15	0.03231429
2020-12-16	0.02887143
2020-12-17	0.02792857
2020-12-18	0.01730000
2020-12-19	0.01788571
2020-12-20	0.01634286
2020-12-21	0.02175714
2020-12-22	0.02227143
2020-12-23	0.02228571
2020-12-24	0.02497143
2020-12-25	0.02451429
2020-12-26	0.02340000
2020-12-27	0.02382857
2020-12-28	0.02407143
2020-12-29	0.02311429
2020-12-30	0.02690000
2020-12-31	0.02437143
2021-01-01	0.02554286
2021-01-02	0.02700000
2021-01-03	0.02594286
2021-01-04	0.02102857
2021-01-05	0.02042857
2021-01-06	0.01867143
2021-01-07	0.01851429
2021-01-08	0.01801429
2021-01-09	0.01784286
2021-01-10	0.01828571
2021-01-11	0.01845714
2021-01-12	0.01805714
2021-01-13	0.01687143
2021-01-14	0.01798571
2021-01-15	0.02034286
2021-01-16	0.01968571
2021-01-17	0.01924286
2021-01-18	0.01747143
2021-01-19	0.01817143
2021-01-20	0.01958571
2021-01-21	0.01851429
2021-01-22	0.01785714
2021-01-23	0.01712857
2021-01-24	0.01785714

DATE	throughput_rolling_avg
2021-01-25	0.01734286
2021-01-26	0.01977143
2021-01-27	0.01787143
2021-01-28	0.01760000
2021-01-29	0.01624286
2021-01-30	0.01722857
2021-01-31	0.01928571
2021-02-01	0.01960000
2021-02-02	0.01601429
2021-02-03	0.01628571
2021-02-04	0.02231429
2021-02-05	0.02085714
2021-02-06	0.02062857
2021-02-07	0.01970000
2021-02-08	0.02051429
2021-02-09	0.02105714
2021-02-10	0.02074286
2021-02-11	0.01535714
2021-02-12	0.01562857
2021-02-13	0.01995714
2021-02-14	0.01897143
2021-02-15	0.02094286



7-day rolling average

- Throughput rolling averages showed fluctuations over time, with notable peaks (4th and 11th Dec 2020) and troughs (18-20 Dec 2020). This information can guide the identification of periods of increased or decreased operational efficiency.
- The high variability in the initial days and subsequent stabilization of the throughput can be attributed to the rolling average calculation.
- The 7-day rolling average provides a smoothed representation of throughput trends, minimizing the impact of daily fluctuations.
- The choice between daily metric and rolling average depends on the context and the desired level of granularity. If rapid changes need attention, the daily metric is more suitable; otherwise, the rolling average provides a more stable trend.

Language Share Analysis:

```
WITH lang_count AS (
    SELECT
        LANGUAGE,
        COUNT(*) AS num_events
    FROM
        job_data
    WHERE
        ds BETWEEN DATE_SUB(
            (
                SELECT
                    MAX(ds)
                FROM
                    job_data
            ),
            INTERVAL 30 DAY
        )
        AND (
            SELECT
                DATE(MAX(ds))
            FROM
                job_data
        )
    GROUP BY
        LANGUAGE
)
SELECT
    LANGUAGE,
    num_events,
    ROUND(
        (num_events * 100.0) / (
            SELECT
                SUM(num_events)
            FROM
                lang_count
        ),
        2
    ) AS percentage
FROM
    lang_count
ORDER BY
    percentage;
```

LANGUAGE	num_events	percentage
Hindi	6	13.33
Persian	7	15.56
Italian	8	17.78
English	8	17.78
French	8	17.78
Arabic	8	17.78

- The language distribution was relatively balanced, with key languages contributing equally to the total events. This insight can inform decisions on multilingual support and content creation strategies.
- All languages have an equal percentage share of 17.78%, except Persian, which has a slightly lower percentage of 15.56% and Hindi, which has an even lower percentage of 13.33%.

- The balanced distribution of events across multiple languages highlights the importance of maintaining support for various languages.

Duplicate Rows Detection:

```
# Inserting duplicate data
INSERT INTO job_data
  (ds, job_id, actor_id, event, language, time_spent, org)
VALUES
  ('2020-12-01', 27, 1012, 'skip', 'English', 67, 'C'),
  ('2020-12-01', 27, 1012, 'skip', 'English', 67, 'C'),
  ('2020-12-03', 26, 1032, 'skip', 'Persian', 70, 'B'),
  ('2020-12-03', 26, 1032, 'skip', 'Persian', 70, 'B'),
  ('2020-12-05', 25, 1033, 'skip', 'English', 15, 'D'),
  ('2020-12-05', 25, 1033, 'skip', 'English', 15, 'D'),
  ('2020-12-10', 26, 1014, 'skip', 'English', 63, 'A'),
  ('2020-12-10', 26, 1014, 'skip', 'English', 63, 'A'),
  ('2020-12-19', 30, 1035, 'skip', 'English', 52, 'D'),
  ('2020-12-19', 30, 1035, 'skip', 'English', 52, 'D'),
  ('2020-12-25', 31, 1046, 'skip', 'English', 77, 'C'),
  ('2020-12-25', 31, 1046, 'skip', 'English', 77, 'C'),
  ('2020-12-29', 38, 1033, 'skip', 'Persian', 40, 'B'),
  ('2020-12-29', 38, 1033, 'skip', 'Persian', 40, 'B');
```

```
# Detecting Duplicate Data
SELECT
    *,
    COUNT(*) AS no_of_records
FROM
    job_data
GROUP BY
    ds,
    job_id,
    actor_id,
    event,
    language,
    time_spent,
    org
HAVING
    COUNT(*) > 1
ORDER BY
    ds;
```

ds	job_id	actor_id	event	language	time_spent	org	no_of
2020-12-01	27	1012	skip	English	67	C	3
2020-12-03	26	1032	skip	Persian	70	B	3
2020-12-05	25	1033	skip	English	15	D	3
2020-12-10	26	1014	skip	English	63	A	3
2020-12-19	30	1035	skip	English	52	D	2
2020-12-25	31	1046	skip	English	77	C	2
2020-12-29	38	1033	skip	Persian	40	B	3

Identifying and addressing duplicate rows is crucial for data quality and integrity. Duplicate rows could potentially lead to inaccuracies in analysis and decision-making. The presence of duplicates emphasizes the importance of data-cleaning processes to ensure the reliability of analytical results.

Results

Through this project, we have achieved a comprehensive understanding of the operational dynamics within the company. The insights derived from the analysis offer valuable information for decision-making processes across departments. Specifically, the identification of patterns in job reviews, throughput trends, and language distribution can contribute to optimizing operational efficiency, improving user experience, and guiding strategic decisions.

The project's outcomes provide actionable insights that can be used to refine operations, allocate resources more effectively, and enhance overall business performance.