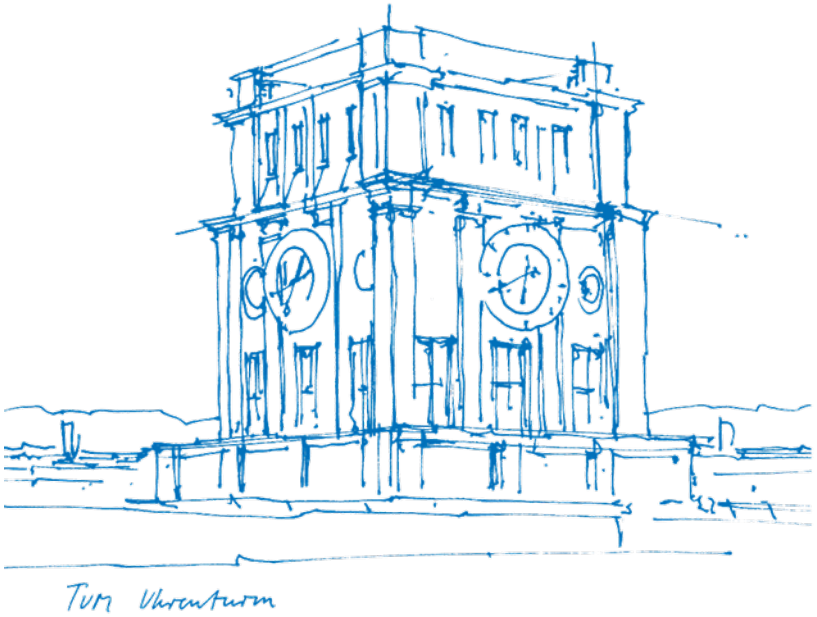
Implementation of a Data Analytics System and Generation of Insights for a B2B SaaS Product

Simon Kreuzer
Technical University of Munich
Department of Informatics
Chair of Decentralized Systems Engineering
Munich, April 27, 2022



Introduction

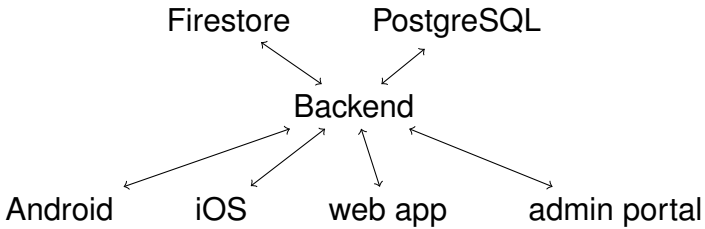
Main goal: Improve the product

- Software products generate humongous amounts of data.
- Big data for processing such data sets: Combine different data sources and convert the data lake into valuable information.

For the B2B SaaS product *deskbird* this means:

- Compare solutions to combine data sources and run queries.
- Set up the data analytics system.
- Develop insights to support decision making.

Infrastructure of deskbird



- Firestore database (NoSQL) contains all information for our backend analyses.
- Focus on workspace booking.

company ⟶ office ⟶ floor (group) ⟶ area (zone) ⟶ workplace

Data warehouse

Connects various heterogeneous sources in one place to run complex queries for analyses.

- Cloud based warehouse for a central instance.
- Hosted warehouse for simple setup and many extra features and integrations.
- Self-hosted and open source warehouses for more flexibility.

⇒ Google BigQuery as cloud based and hosted solution for deskbird.

Analytics system setup

Mirror data to BigQuery

- Mirror collections to BigQuery with the provided extension *Stream Collections to BigQuery*.
- BigQuery tables with columns `timestamp`, `event_id`, `document_name`, `operation`, `document_id` and `data`.
- The `data` column contains the firestore document as JSON string.

Extract fields from JSON string and adjust data types using views.

```
1 SELECT
2   JSON_VALUE(DATA, '$.id') AS id,
3   JSON_VALUE(DATA, '$.name') AS name,
4   SAFE_CAST(JSON_VALUE(DATA, '$.allowsOfficePlanning') AS BOOLEAN) AS officePlanning,
5   TIMESTAMP_MILLIS(SAFE_CAST(JSON_VALUE(DATA, '$.createdAt') AS INT64)) AS creationTime
6 FROM
7   `deskbird-bbe72.firestore_mirror.businessCompanies_raw_latest`
```

Analytics system setup

id	name	officePlanning	creationTime
⋮	⋮	⋮	⋮

Views used as base for further SQL query development.

- `businessCompanies`
- `users`
- `bookings`
- `internalworkspaces`
- `zones`

Development of historical overview of users and bookings

Goal: Historical overview of users and booking per company and both per month and day with the following metrics:

- *Amount of users*: Registered users within the given month or day.
- *Amount of active users*: Users with at least one booking within the given month or day.
- *Amount of bookings*: Bookings within the given month or day.

Development of historical overview of users and bookings

Amount of newly registered users per company and month.

```
1 SELECT
2   companyId,
3   DATE_TRUNC(EXTRACT(DATE FROM creationTime), MONTH) AS month,
4   COUNT(DISTINCT id) AS numberUsers
5 FROM
6   `deskbird-bbe72.firestore_mirror.users`
7 GROUP BY
8   companyId,
9   month
```

Development of historical overview of users and bookings

Amount of active users and bookings per company and month.

```
1  SELECT
2    companyId,
3    DATE_TRUNC(EXTRACT(DATE FROM creationTime), MONTH) AS month,
4    COUNT(DISTINCT userId) AS numberActiveUsers,
5    COUNT(DISTINCT id) AS numberBookings
6  FROM
7    `deskbird-bbe72.firestore_mirror.bookings`
8  GROUP BY
9    companyId,
10   month
```

Development of historical overview of users and bookings

Month array from creation date of the company to the current month.

```
1  SELECT
2    companies.id,
3    companies.name,
4    companies.status,
5    month
6  FROM
7    `deskbird-bbe72.firestore_mirror.businessCompanies` companies,
8    UNNEST(GENERATE_DATE_ARRAY(
9      DATE_TRUNC(EXTRACT(DATE FROM companies.creationTime), MONTH),
10     CURRENT_DATE(),
11     INTERVAL 1 MONTH
12   )) month
```

Development of historical overview of users and bookings

Make the number of users cumulative.

```
1 SUM(numberUsers) OVER (
2   PARTITION BY
3     companies.id
4   ORDER BY
5     month ASC)
```

Left join all metrics to the companies.

⇒ All companies with all months they are existing in the database are included, even if a month or company has no users or bookings.

Development of historical overview of users and bookings

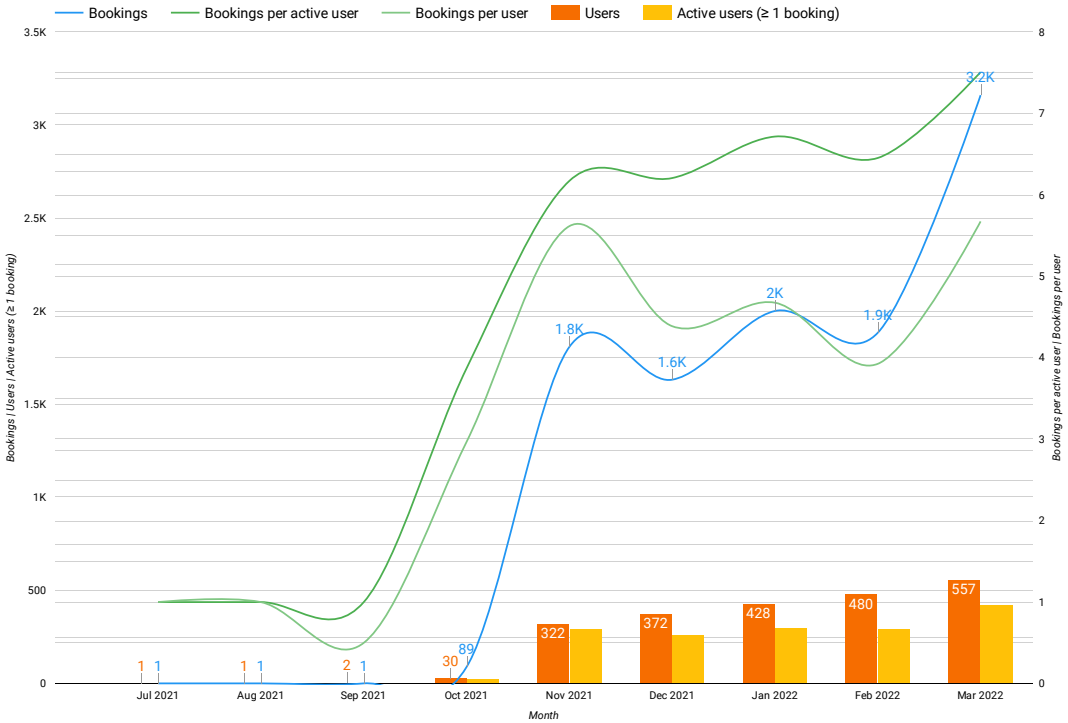
Results of the full query:

id	name	status	month	numberUsers	numberActiveUsers	numberBookings
⋮	⋮	⋮	⋮	⋮	⋮	⋮

Additional metrics by combining the existing ones:

- `numberBookings / numberUsers`
- `numberBookings / numberActiveUsers`

Historical overview of users and bookings



Google DataStudio for visualization

- Random, active German customer of deskbird.
- Testing phase in the beginning.
- Drop in December because of the holidays.
- Strong increase of bookings and active users in March because of the repeal of the home office obligation.

Conclusion

Achievements of the data analytics setup and developed insights:

- Better customer approaching
- Feature improvement

Future work

- Continue with more detailed analyses
- Migration from Firestore to SQL databases $\rightarrow$ Re-setup of mirroring
- Track user behavior by monitoring the clients

Bibliography I

[1] Y. Li and S. Manoharan, “A performance comparison of sql and nosql databases,” in *2013 IEEE Pacific Rim Conference on Communications, Computers and Signal Processing (PACRIM)*, Aug. 2013, pp. 15–19. DOI: 10.1109/PACRIM.2013.6625441.

[2] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, T. Chandra, A. Fikes, and R. Gruber, “Bigtable: A distributed storage system for structured data (awarded best paper!)” In *7th Symposium on Operating Systems Design and Implementation (OSDI ’06), November 6-8, Seattle, WA, USA*, B. N. Bershad and J. C. Mogul, Eds., USENIX Association, 2006, pp. 205–218.

[3] K. Chodorow and M. Dirolf, *MongoDB - The Definitive Guide: Powerful and Scalable Data Storage*. O’Reilly, 2010, ISBN: 978-1-449-38156-1.

[4] J. C. Anderson, J. Lehnardt, and N. Slater, *CouchDB - The Definitive Guide: Time to Relax*. O’Reilly, 2010, ISBN: 978-0-596-15589-6.

[5] R. K. Chawda and G. Thakur, “Big data and advanced analytics tools,” in *2016 Symposium on Colossal Data Analysis and Networking (CDAN)*, Mar. 2016, pp. 1–8. DOI: 10.1109/CDAN.2016.7570890.

Bibliography II

[6] R. Patgiri and A. Ahmed, "Big data: The v's of the game changer paradigm,", Dec. 2016. DOI: 10.1109/HPCC-SmartCity-DSS.2016.0014.

[7] *Cloud firestore*, https://firebase.google.com/docs/firestore, Mar. 2022.

[8] Firebase, *Stream collections to bigquery*, https://firebase.google.com/products/extensions/firebase-firestore-bigquery-export, Mar. 2022.

[9] *Cloud sql for postgresql documentation*, https://cloud.google.com/sql/docs/postgres, Mar. 2022.

[10] *Extract tranform load (etl)*, https://databricks.com/glossary/extract-transform-load, Mar. 2022.

[11] M. Golfarelli and S. Rizzi, "From star schemas to big data: 20+ years of data warehouse research," in *A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years*, S. Flesca, S. Greco, E. Masciari, and D. Saccà, Eds. Cham: Springer International Publishing, 2018, pp. 93–107, ISBN: 978-3-319-61893-7. DOI: 10.1007/978-3-319-61893-7_6.

[12] S. Fernandes and J. Bernardino, "What is bigquery?" In *Proceedings of the 19th International Database Engineering & Applications Symposium*, ser. IDEAS '15, Yokohama, Japan: Association for Computing Machinery, 2015, pp. 202–203, ISBN: 9781450334143. DOI: 10.1145/2790755.2790797.

[13] I. Pandis, "The evolution of amazon redshift," *Proc. VLDB Endow.*, vol. 14, no. 12, pp. 3162–3163, 2021.

Bibliography III

[14] N. Shakhovska, N. Boyko, and P. Pukach, "The information model of cloud data warehouses," in *Advances in Intelligent Systems and Computing III - Selected Papers from the International Conference on Computer Science and Information Technologies, CSIT 2018, September 11-14, Lviv, Ukraine*, N. Shakhovska and M. O. Medykovskyy, Eds., ser. Advances in Intelligent Systems and Computing, vol. 871, Springer, 2018, pp. 182–191. DOI: 10.1007/978-3-030-01069-0_13.

[15] *Data warehouse architecture: Traditional vs. cloud*, https://panoply.io/data-warehouse-guide/data-warehouse-architecture-traditional-vs-cloud/, Mar. 2022.

[16] N. Samuel, *Sql vs nosql databases: 5 critical differences*, https://hevodata.com/learn/sql-vs-nosql-databases-5-critical-differences/, Mar. 2021.

[17] *Bigquery documentation*, https://cloud.google.com/bigquery/docs/, Mar. 2022.

[18] F. Yang, E. Tschetter, X. Léauté, N. Ray, G. Merlino, and D. Ganguli, "Druid: A real-time analytical data store," in *International Conference on Management of Data, SIGMOD 2014, Snowbird, UT, USA, June 22-27, 2014*, C. E. Dyreson, F. Li, and M. T. Özsu, Eds., ACM, 2014, pp. 157–168. DOI: 10.1145/2588555.2595631.

[19] G. s. Bhathal and A. S. Dhiman, "Big data solution: Improvised distributions framework of hadoop," in *2018 Second International Conference on Intelligent Computing and Control Systems (ICICCS)*, Jun. 2018, pp. 35–38. DOI: 10.1109/ICCONS.2018.8663142.

Bibliography IV

[20] Z. Lyu, H. H. Zhang, G. Xiong, G. Guo, H. Wang, J. Chen, A. Praveen, Y. Yang, X. Gao, A. Wang, W. Lin, A. Agrawal, J. Yang, H. Wu, X. Li, F. Guo, J. Wu, J. Zhang, and V. Raghavan, "Greenplum: A hybrid database for transactional and analytical workloads," in *SIGMOD '21: International Conference on Management of Data, Virtual Event, China, June 20-25, 2021*, G. Li, Z. Li, S. Idreos, and D. Srivastava, Eds., ACM, 2021, pp. 2530–2542. DOI: 10.1145/3448016.3457562.

[21] A. Abelló and O. Romero, "Online analytical processing," in *Encyclopedia of Database Systems*, L. Liu and M. T. Özsu, Eds. New York, NY: Springer New York, 2018, pp. 2558–2563, ISBN: 978-1-4614-8265-9. DOI: 10.1007/978-1-4614-8265-9_252.

[22] N. Pendse, R. Creeth, and B. (Firm), *The OLAP Report: Succeeding with On-line Analytical Processing*. Business Intelligence, 1995, ISBN: 9781898085218.

[23] B. Hüsemann, J. Lechtenbörger, and G. Vossen, "Conceptual data warehouse design,", Jul. 2000.

[24] S. Huber and N. Litzel, *Was ist ein data warehouse?* https://www.bigdata-insider.de/was-ist-ein-data-warehouse-a-606701/, May 2017.

[25] E. Zagan and M. Danubianu, "Cloud data lake: The new trend of data storage," in *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, Jun. 2021, pp. 1–4. DOI: 10.1109/HORA52670.2021.9461293.

Bibliography V

[26] G. Snipes, "Google data studio," *Journal of Librarianship and Scholarly Communication*, vol. 6, no. 1, 2018. DOI: 10.7710/2162-3309.2214.

[27] B. Santos, F. Sério, S. Abrantes, F. Sá, J. Loureiro, C. Wanzeller, and P. Martins, "Open source business intelligence tools: Metabase and redash," in *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, IC3K 2019, Volume 1: KDIR, Vienna, Austria, September 17-19, 2019*, A. L. N. Fred and J. Filipe, Eds., ScitePress, 2019, pp. 467–474. DOI: 10.5220/0008351704670474.

[28] K. Gudfinnsson and M. Strand, "Challenges with bi adoption in smes," in *2017 8th International Conference on Information, Intelligence, Systems & Applications (IISA)*, Aug. 2017, pp. 1–6. DOI: 10.1109/IISA.2017.8316407.