
GROFERS


Data
Hub

About me

Shubham Gupta / shubhamg931.tech

Data Engineer @ Grofers

 in/shubhamg931  shubhamg931

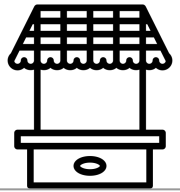
Agenda

1. Grofers
2. Data at Grofers
3. Need for a Data Catalog
4. Why Datahub?

About Grofers

27+
Cities

5,000+
Vendors



40,000+
Daily Orders

Data at Grofers

Central Data Engineering Team

10 Engineers

2 Data Scientists

2 Data Analysts

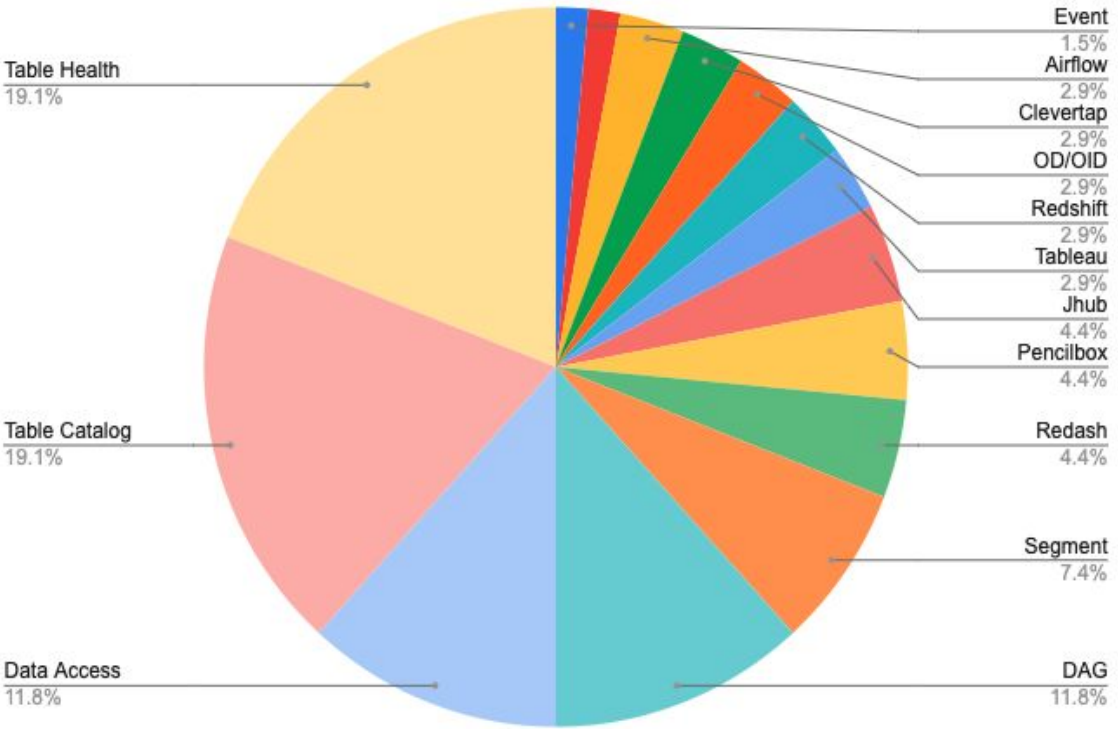
2 Product managers

Data Analysts embedded in Business Teams

~50 Data Analysts

Need for a Data Catalog

Frequently Asked Questions



Data discovery is difficult

1. **Number of Tables, Variety of ETLs**

   Multiple people reaching out to centralized data team for questions about more than 2000 tables and ETLs

2. **Transfer of Knowledge**

   People who created the tables leaves the organization at some point, and all the knowledge goes with them

3. **Evolving nature of Data**

   Tables' structure changes over time and needs to be dynamically updated

Problems it causes

1. **Inefficiency**
   People have to wait for their question to be answered

2. **Duplicacy**
   Existing tables do not get explored and new tables get created with minor changes

3. **Distrust**
   Not all questions get answered and people do not know about table health as well which causes lack of trust in data

Why Datahub?

Features

	Datahub	Amundsen	Atlas
Search	✔️	🔥	✔️
Data Lineage	🔥	❌	✔️
Business Glossary	🔥	✔️	✔️
Roadmap	🔥	✔️	❌

S.No	Features	Reach	Impact	Feature need (Reach * Impact)	DataHub		Amundsen		Atlas	
					Description	Score	Description	Score	Description	Score
1	Data Lineage	8	8	64	Supports lineage graphs, and future plans of column based lineage	5	Not natively supported apart from table-dashboard links for redash dashboards. Will probably make use of airflow/atlas capabilities, but in very early discussion stage	2	Support lineage graphs and classification propagation	5
2	Data Marketplace	5	2	10	Not available	0	Preview dataset option is available	5	Not supported	0
3	Metadata Management	8	8	64	Publish to kafka topic and it syncs in neo4j and db and es	8	We can sync metadata at scheduled times, and also push desired metadata at end of each jobs	8	Support to create entities/attributes required for metadata	8
4	Search	8	8	64	Available. New fields can be made searchable as well	8	Users can search for data sets, columns tags	8	Available. Various APIs available to search on different factors	8
5	Tagging	5	3	15	Available for users, and can create for datasets easily	8	We can tag data sets	8	Supported	8
6	Business Glossary	8	8	64	Descriptions, owners, document links	8	Can have descriptions	5	Supported	5
7	Automated updation	5	8	40	Can drive it with airflow	5	Can drive this with airflow, and push jobs with every dag	8	Can drive it with Airflow	5
8	User/User Groups	5	5	25	Can have users currently, but cant group them or do selective permissions	5	Can have users currently, but cant group them or do selective permissions	3	Support to create entities/attributes required for metadata	3
9	Social Features	8	3	24	Can see others' profile and their owned datasets	5	Can see others' profiles and their bookmarks	5	Not available	0
10	Dashboard tool connectors(Redash)	5	3	15	No, but can build entity and need to find a way to dump	3	Connector for redash dashboards onboarding	5	Not available	0
11	Viz tool connector(Tableau/powerbi/superset)	5	3	15	No, but can build entity and need to find a way to dump	3	Available for superset	3	Not available	0
12	Warehouse/lake Support(Redshift, Postgres, hive)	5	3	15	Most sql	5	Redshift, postgres	5	Not available	0
13	Ease of Management	3	8	24	Neo4j, kafka, schema registry, es, services and java	5	Neo4j, es, and services, and python	8	Hadoop, HBase, Solr, Janusgraph and Java. Admin view UI is there, will need to develop proper user view	1
Total Score (Sum (Feature need * Score) / 100)					**27.06**		**25.44**		**20.83**	

Maturity

1. **Documentation**

   Most properly documented tool

2. **Community**

   Core team very active on slack and new features are being actively worked upon.

Thanks!

Special thanks to active efforts of the Datahub team who helped us build our POC, we were able to finalize Datahub as our Data cataloging tool :)