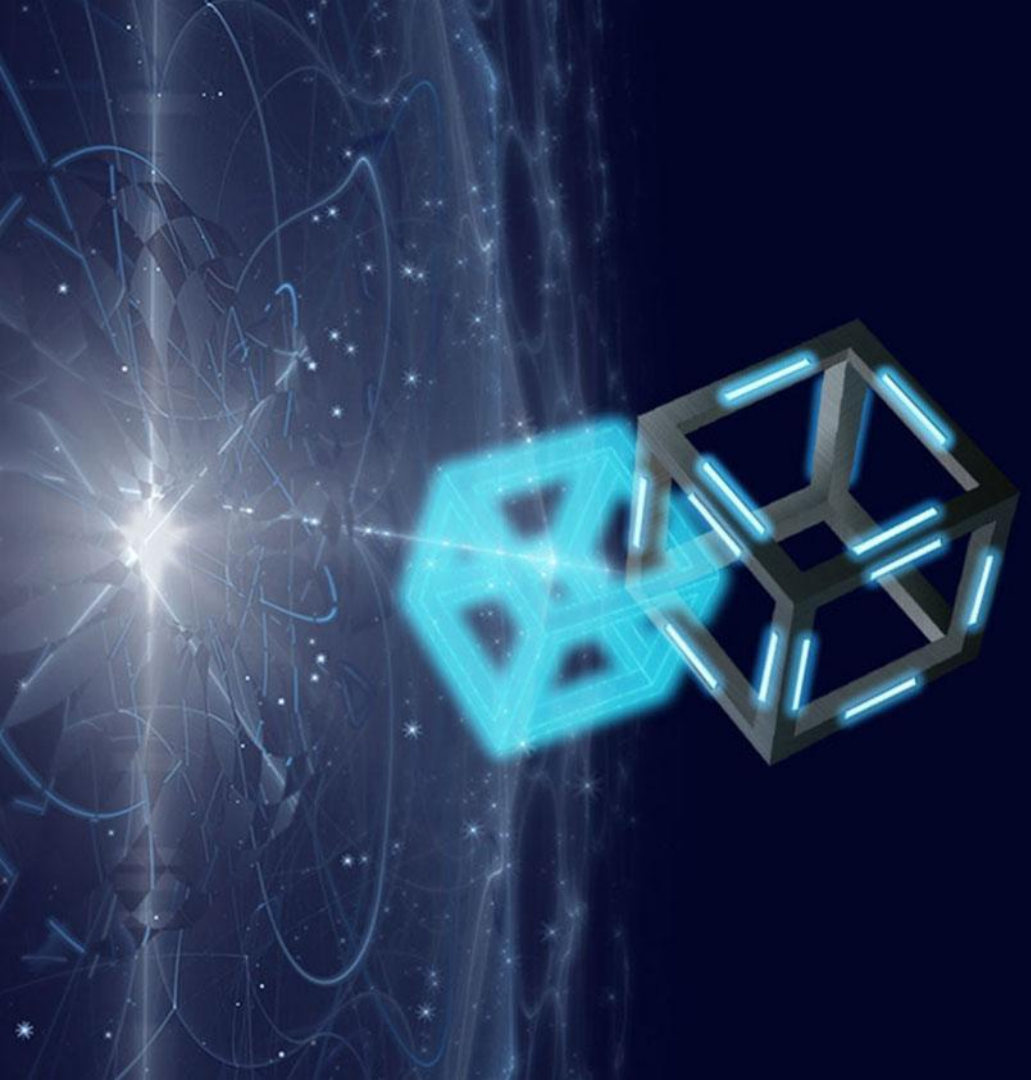
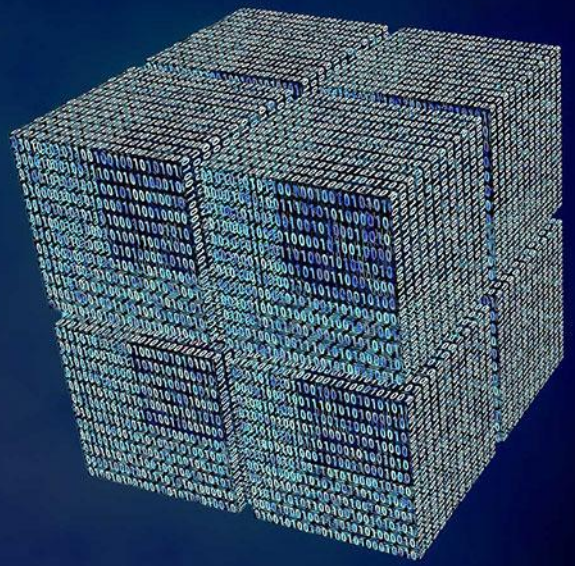


Supercharged Industry Database

Revolutionize conventional database to accommodate Big Data application.

Zhikun Chen, Wan Ting Hsu, Ariel Xie, Zhen Liu

Table of Content



Design Overview

Fundamental research for equity purposes are being revolutionized. **90%** of the IBs are using SQL or Excel to store and analyze data, which is not enough

Our Supercharged Industry database would be able to increase efficiency (speed and cost) and real time Big Data (social media) for analyst during fundamental research. It also increases data security and functionality.

Propose Design

Our Supercharged Industry Databases can enhance the performance when dealing with unstructured data and large databases. Upgrade the storage, and speed. Build on top of NoSQL database in order to solve SQL database weakness and expand the alternative data storage.
Our Product can process big-data workload and to make future data predictions. That could help us achieve both high-quality algorithms and high speed.
Furthermore, we can utilize kibana and Tableau for automated data visualization.

Product Function Overview

User characteristics	Investment Banking Analyst; Competitors in this industry; National organizations
General constraints	The collision and evolution of old (SQL) and new ideas (NoSQL)
Impact & Benefit	Data sharing. Operational efficiency will be improved.Improves data integrity and security. Data consistency and maintainability can be guaranteed
General Assumption & Dependencies	Higher demand for NoSQL will be higher requirements.

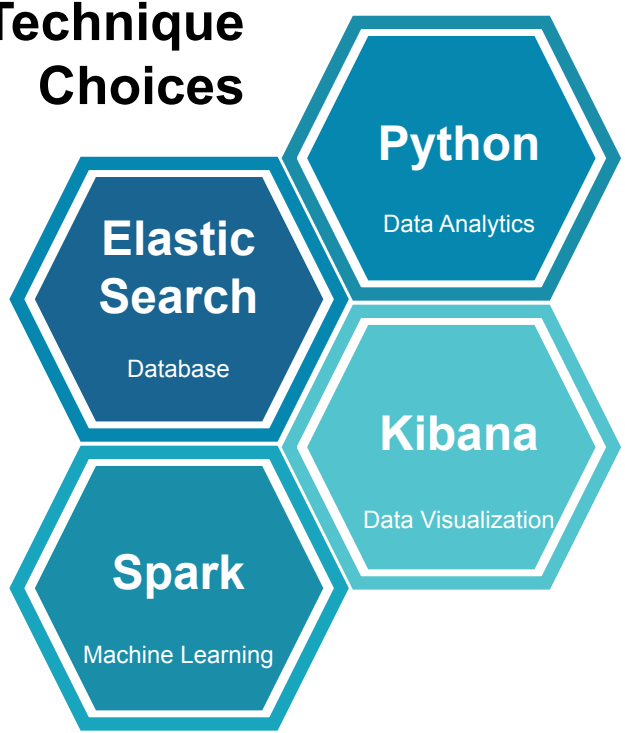
Data Source Specification

Inputs & Outputs
Input: data vendors and data suppliers
Output: Industry database

Data Sources
Wind & Baiinfo

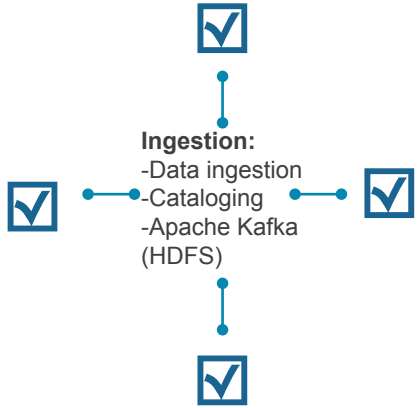
Procurement Details
Searching Prebaked Anode Dataset through Desktop Application such as Wind, Baiinfo, Hithink Flush and so on.
Discussion with experts in Prebaked Anode industry
Using NLP and Neural Networks to analyze satellite images, etc.

Technique Choices

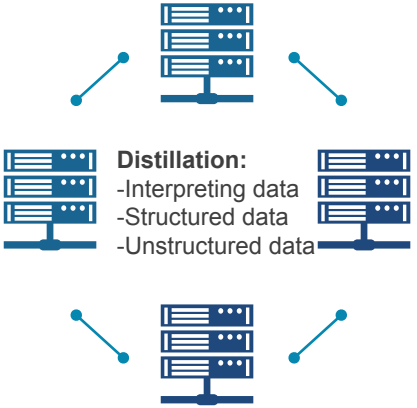


Data Lake Architecture

Sources: Raw Data
FIle data, relational data, streaming data

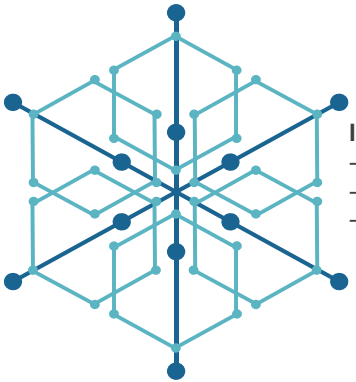


Ingestion:
-Data ingestion
-Cataloging
-Apache Kafka
(HDFS)



Distillation:
-Interpreting data
-Structured data
-Unstructured data

Processing:
-ElasticSearch
-Python
-Machine Learning
Tool: Spark
Business Logic

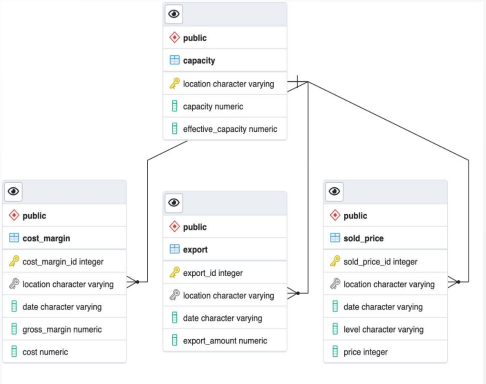


Insights:
-Kibana
-Metabase
-Data Visualization

Action: Business System
Data base, visualization reports, external data

Relational Database

Create table and database



1. Analyzing raw data
2. Normalization & Design ERD for database
3. ELT Process
4. Insert Data

#ELT Example

```
cost_margin = cost_margin.drop_duplicates()
cost_margin.insert(0, 'cost_margin_id', range(20000, 20000 +
len(cost_margin)))
cost_margin = pd.DataFrame(cost_margin)
cost_margin = cost_margin.rename(columns={"Province": "location",
"Time": "date", "Cost": "cost","Gross margin": "gross_margin"})
cost_margin
```

Query

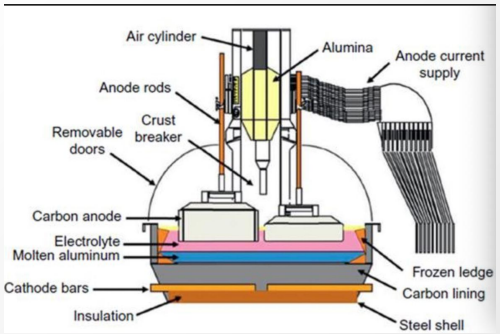
We want to know the highest and lowest sold price in Shandong within the period from May 2020 to May 2021, and that month's cost and gross margin.

```
#1. find the month that has the highest, lowest sold price, and that month's cost and gross margin in Shandong
queryCmd = """

select t1."date",t1."location", t1.high_price, t1.low_price, t2.cost, t2.gross_margin
    from(
    select "date","location"
      ,sum(case when "level" = 'high' then price else 0 end) as high_price
      ,sum(case when "level" = 'low' then price else 0 end) as low_price
    from Sold_price
    where "location" = 'Shandong'
    group by "date","location"
) as t1
left join(
    select "date","location",sum(cost) as cost,sum(gross_margin) as gross_margin
    from Cost_Margin
    group by "date","location"
) as t2
on t2."date" = t1."date" and t2."location" = t1."location"

"""
cur.execute(queryCmd)

row=cur.fetchone()
while row is not None:
    print("Month:", row[0], ",  Location: ", row[1], ", Cost: ", row[4],", Gross Margin: ", row[5],)
    row = cur.fetchone()
```

```
Month: 9/1/20 ,  Location:  Shandong , Cost:  14671.67 , Gross Margin:  159.26
Month: 8/1/20 ,  Location:  Shandong , Cost:  14527.67 , Gross Margin:  40.0
```

Managing Unstructured Data

01 Image



02 Image exif data

```
{'size': (591, 393),
 'image_format': 'JPEG',
 'image_mime': 'image/jpeg',
 'datetime': '2022-04-18 21:42:12.465817',
 'uuid': 'd3380bec-a54a-42d5-8fd6-7cbdd278d5f9',
 'make': 'Camera Unknown',
 'model': 'Camera Unknown',
 'software': 'Unknown Software',
 'x_res': None,
 'y_res': None,
 'name': 'Prebaked-anode.jpg'}
```

03 Image array

```
Output exceeds the size limit.
[[[255, 255, 255],
  [255, 255, 255],
  [255, 255, 255],
  [255, 255, 255],
  [255, 255, 255],
  [255, 255, 255],
  [255, 255, 255],
  [255, 255, 255],
  [255, 255, 255],
  [255, 255, 255],
```

04 Base64 encode; Save the image

```
img_base64 = base64.b64encode( bytes(img_str, "utf-8") )
```

Data Policies & Cost Implications

Process

- Data collection from public data, industry data, social media, news etc.
- Data cleaning and data integration.
- Set up data searching glossary (NLP dictionary)
- Set up data visualization

Mission

- Create a more efficient way to search and analyze data to combat the most time/cost consuming task in IB industry - fundamental research reports
- Increase data storage within the company database and minimize costs at the same time .

Technologies

- Use Elasticsearch-DSL and Elasticsearch package on Python to load industry data we retrieved from data vendors.
- Create connection between elasticsearch database and kibana and use kibana for data visualizations.
- Conduct machine learning for unstructured data to predict valuable information.

Cost Implications

- Date quality will lead to accurate analysis, customer relations, and business decisions, so in order to maintain the quality of data, there is necessary cost implications, such as basic fee, data supplier fee, market data fee, server fee

Evaluation Criteria

Quantitative success

- 50% of the analysts in the Prebaked Anode industry will use our database to complete their reports.
- Prediction accuracy using Spark Package

Qualitative success

- Getting positive feedbacks from surveys through clients
- Frequent citation by experts in Prebaked Anode industry

Data Governance: This program will define the master data models, detail the retention policies for data, and define roles and responsibilities for data authoring, data curation, and access. Such as admin-level access, editor access, viewer access, and public access.

Conclusion & Recommendations

1

Recommendation

New-gen Market Research

Traditional market research can only get you so far. The majority of the data is not able to share and search. To see the complete picture of Supercharged industry data. Leverage the industry data to level up the company's strategy and discover more about your consumers.

Conclusion

Fast, Broad, and Experts

Industry Database

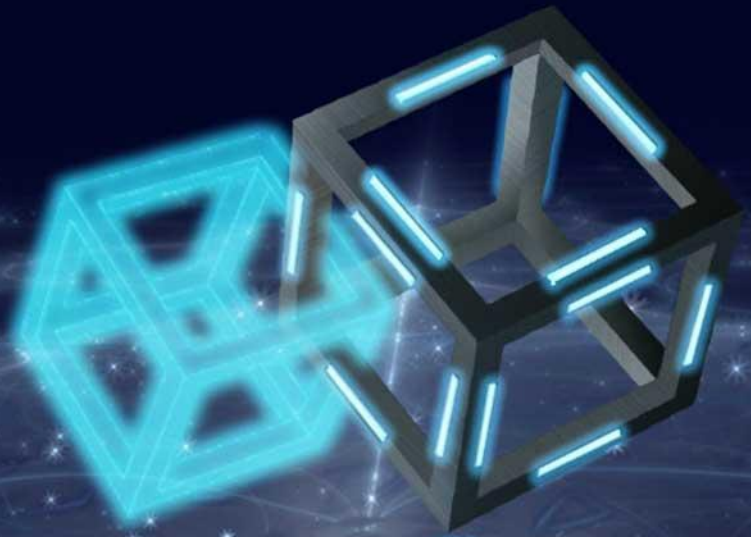
- Receive results in seconds rather than weeks
- Go beyond focus groups and look at the largest pool of online discussion
- Leverage our team of experts to level-up company's work ;)

2

Recommendation

Spot Trend Cross Industry

With full historical data back to 2020, you can monitor trends in every season and month. Search on the database could help you identify the opportunity. Look outside your business and compare it with the outside data trend.



THANK YOU