royjafari / optimizing-big-data-problem-statement Public

Code Issues Pull requests Actions Projects Wiki Security Insights Se

main

optimizing-big-data-problem-statement / ch5 / Solutions.MD

royjafari Update Solutions.MD History

1 contributor

75 lines (57 sloc) 6.69 KB

Solutions to the Excerciese and Studies in Chapter 5

Exercise 1 – Which problem statement is concise?

Answer: The third one - "*We have found a repeatable flaw in the Twitter Intelligence system. The codes downstream to the tweet digestion will fail if there is a gap in data coming from the tweet digestion code; gaps in the data are possible and unavoidable.*" - is a concise problem statement. The first one has unnecessary information, and the second one is a possible solution to the problem.

Exercise 2 – Which problem statement is cross-functional?

Answer: The second one - *" We have found a flaw that can happen again and again in the Twitter Intelligence system. The codes that are run after the tweet digestion fail because they expect all the historical data to come from the tweet digestion code."* is cross-functional. The second one uses many technical words that are largely understood by data engineers.

Exercise 3 – Which problem statement is concise, cross-functional, and knowledge-driven?

Answer: The third one - *“We have found a repeatable flaw in the Twitter Intelligence system. The codes downstream to the tweet digestion will fail if there is a gap in data coming from the tweet digestion code; gaps in the data are possible and unavoidable*.” - is the middle ground between cross-functional and knowledge-driven. The first one is very concise and knowledge-driven, and the second one is too cross-functional.

Exercise 4 – Which problem statement is atomic?

Answer: The three separate problem statements are listed below.

- Subscribers complaining about the unavailability of the services
- Tweet digestion system has failed
- Downstream codes to the tweet digestion system expect full data, and the gap in the data will lead to errors.

Exercise 5 – which problem statements are value centered and measurable?

Answer:

- The first one is both measurable and value-centered. The metric to evaluate the possible solutions is the time it takes for the algorithm to work, and the value is the competitiveness of Google’s search engine usage.
- The second one is only measurable. The metric is how fast the API responds. It is, however, not value-centered.
- The third one is not value centered or measurable.

Exercise 6 - Solving the wrong problem

Answer: The third one is the highest priority. The reason is that in the second stage of the Data Solution Life Cycle, we don’t yet if we have an MVP, so our priority is getting to the point to answer if there is an MVP or not, and the matters of efficiency such as how long a piece of code takes is not as important as matters of effectiveness such as if the data integration is correct.

Exercise 7 – Recognizing the right big-data problem statement

Point of Interest (POI) Recommendation

1. No, it is not cross-functional. We wouldn’t expect all the data professionals to know “Association Rules Analysis“, “A Priority Algorithm” or “Itemset”.

2. Yes. The statement is Knowledge driven. The statement is full of technical terms.
3. No, the statement is not value-centered. It is not obvious why giving recommendations to users will be helpful for the application.
4. Yes, the statement is measurable. The search time is the metric.
5. Stage 5 – Deployment

Ambulance Priority Management

1. Yes, it is cross-functional. There are almost no technical terms in the statement.
2. Yes. The statement is Knowledge driven. You have to know the concept of “data solution” to be able to understand it.
3. No, the statement is not value-centered. It is not obvious why having a severity score can be helpful.
4. Yes, the statement is measurable, but not fully. One measure is the 20-second cut-off point for the calculation of the severity score, but it is not mentioned how the severity score will be evaluated.
5. Stage 2 – Data Acquisition & Integration

Predicting solar energy generation

1. Yes, it is cross-functional. There is not a lot of specialized terminology in the statement. The only technical word is “train” which is a very commonplace concept among all data professionals.
2. Yes. The statement is Knowledge driven. You have to know the concept of “train” and “algorithm” to be able to understand.
3. Yes, the statement is value-centered. The value of improving the algorithm of a self-driving car is evident.
4. Yes, the statement is measurable. If we find a new source of data with a ratio of rare events higher than 1% we know we have been successful.
5. Stage 2 – Data Acquisition & Integration

Self-driving car algorithm

1. No, it is not cross-functional. The terms that are perhaps too technical are “jet stream”, “MVP”, and “processing latency”.
2. Yes. The statement is Knowledge driven. You have to know the concept of “MVP”, “accuracy”, and “predict” to be able to understand.
3. No, the statement is value-centered. The audience might not know why predicting the amount of electricity generation might be useful.
4. Yes, the statement is measurable. Metric: processing latency.
5. Stage 5 – Deployment

Exercise 8 – writing the right problem statement

Case Study: Employee Relationship Management. Theme: Stage 2 Data Acquisition & Integration.

Answer: This is only one of many possible answers.

We want to integrate the following two data sources, but their aggregation level does not match.

- Employees' performance review
- Employees' operational data, for example, their online communications with customers and colleague

The first one's new data objects come quarterly, and the second one's new data objects come hourly. How should we aggregate the data?

Case Study: Dynamic Pricing. Theme: Stage 3 Data Cleaning & Massaging.

Answer: This is only one of many possible answers.

The competition web scraped pricing data has missing values. If the scraping scripts fail due to various reasons, the price will be recorded as NULL. How should we deal with these missing values so the price optimization algorithm is most successful?

Case Study: Credit Card Transaction Fraud Detection. Theme: Stage 5 Deployment

Answer: This is only one of many possible answers.

A new Fraud Detection system can decide if a transaction request is fraudulent or not in 2.5 seconds. Two ways can reduce this to milliseconds territory. 1) Code optimization (a 2-month job for an experienced data engineer) or 2) Purchasing more powerful compute resources from AWS. The first one will incur a one-time cost of $20,000 (the compensation for Data Engineer) and the system will be usable in three months, the second one the system will be usable from tomorrow but our AWS bill increase by $200 monthly. Which strategy should we use?

Give feedback