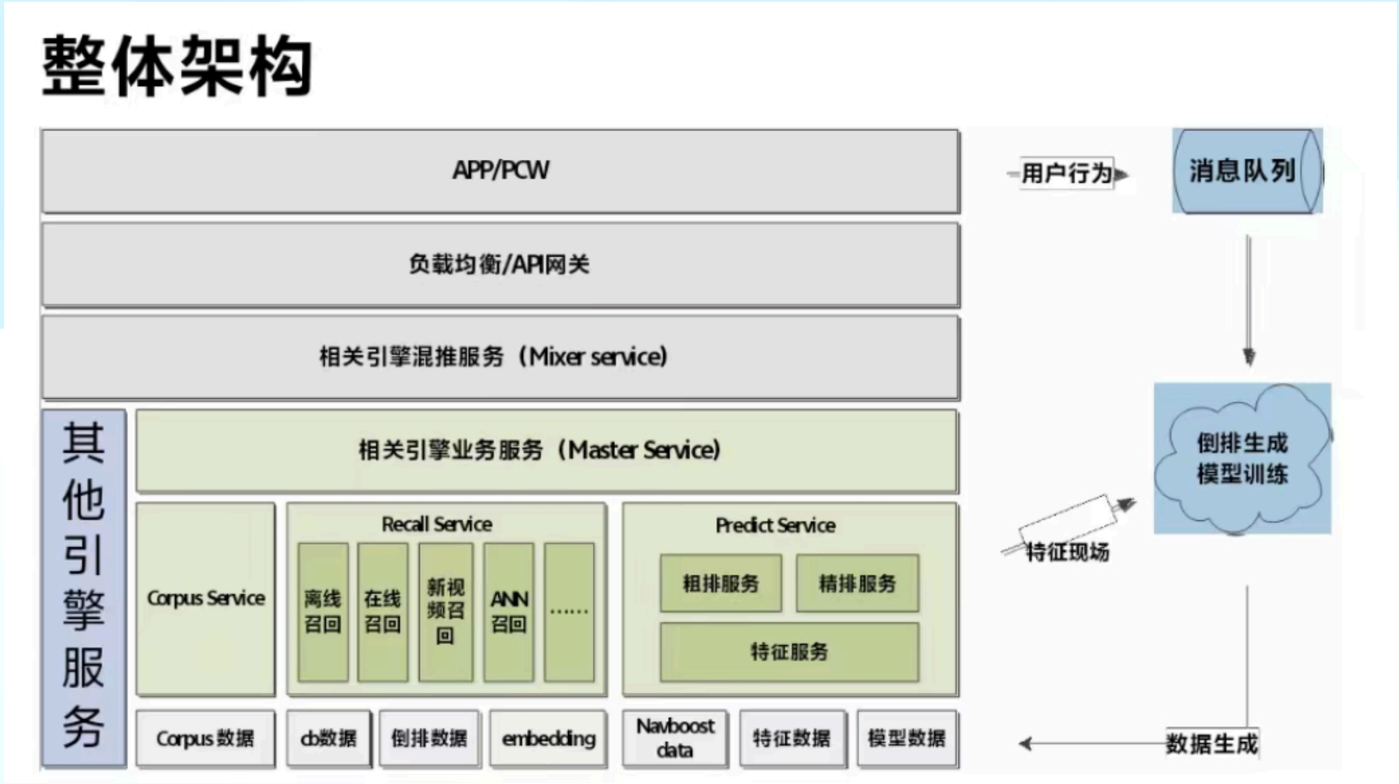
Recommendation system architecture

Recommended system

Engineering and Algorithms

- 

Module interpretation

master service

Module interpretation

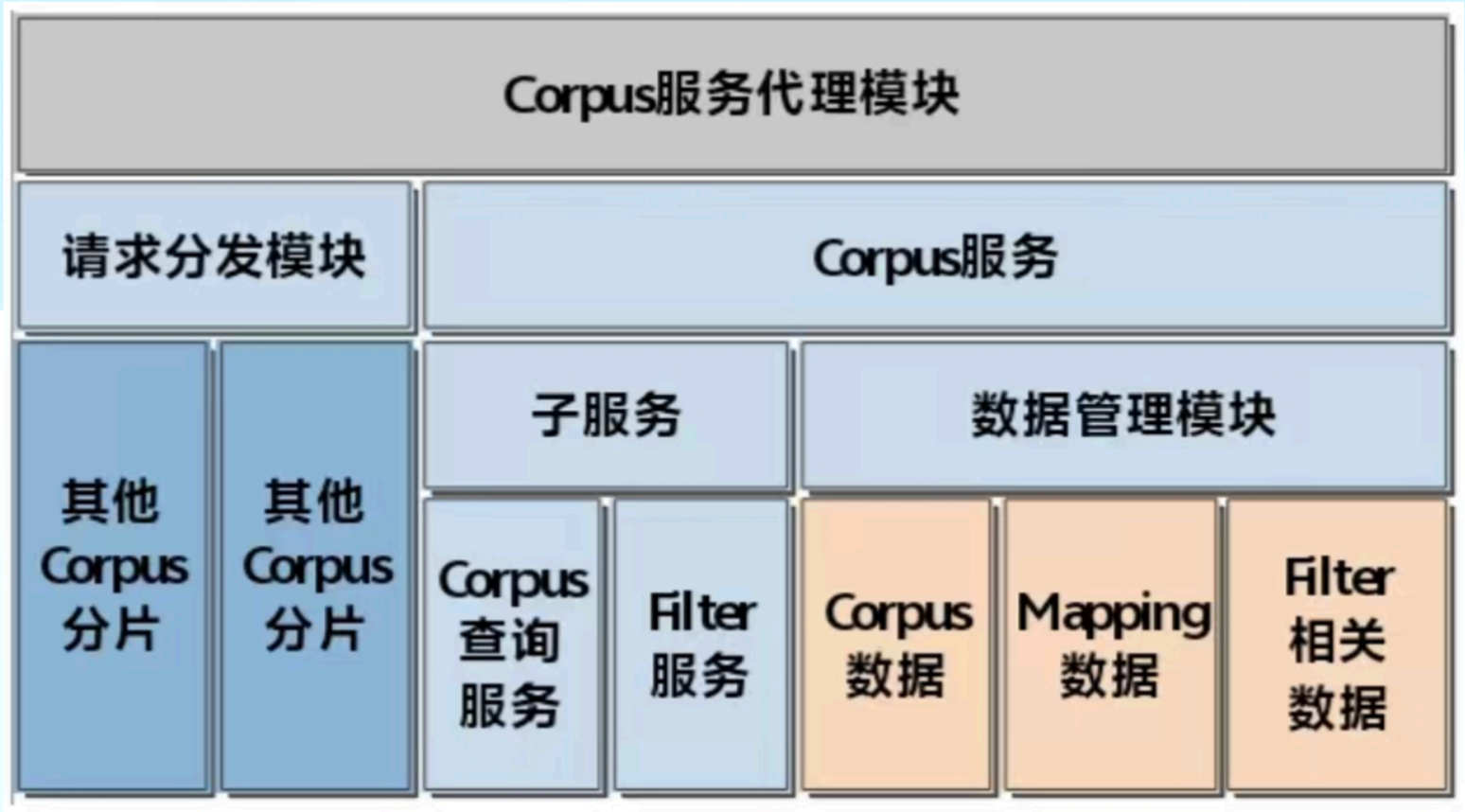
master service

- The Master service is responsible for the main process control logic recommended by related engines and is the logic control center of related engines. This module will follow different logical processes for different requests according to the configuration. This module is only responsible for the overall process of message processing, and the specific processing logic (such as recall logic and sorting logic) is distributed to the corresponding functional modules for processing.

- During a recommendation process, the Master service will call the recall service, corpus service, rough ranking service, and fine ranking service respectively according to the specific configuration information, and will perform diversity control and rerank adjustment on the final results. The Master service is also responsible for loading and managing bucket configuration files, ctl data, etc.

Module interpretation

corpus services



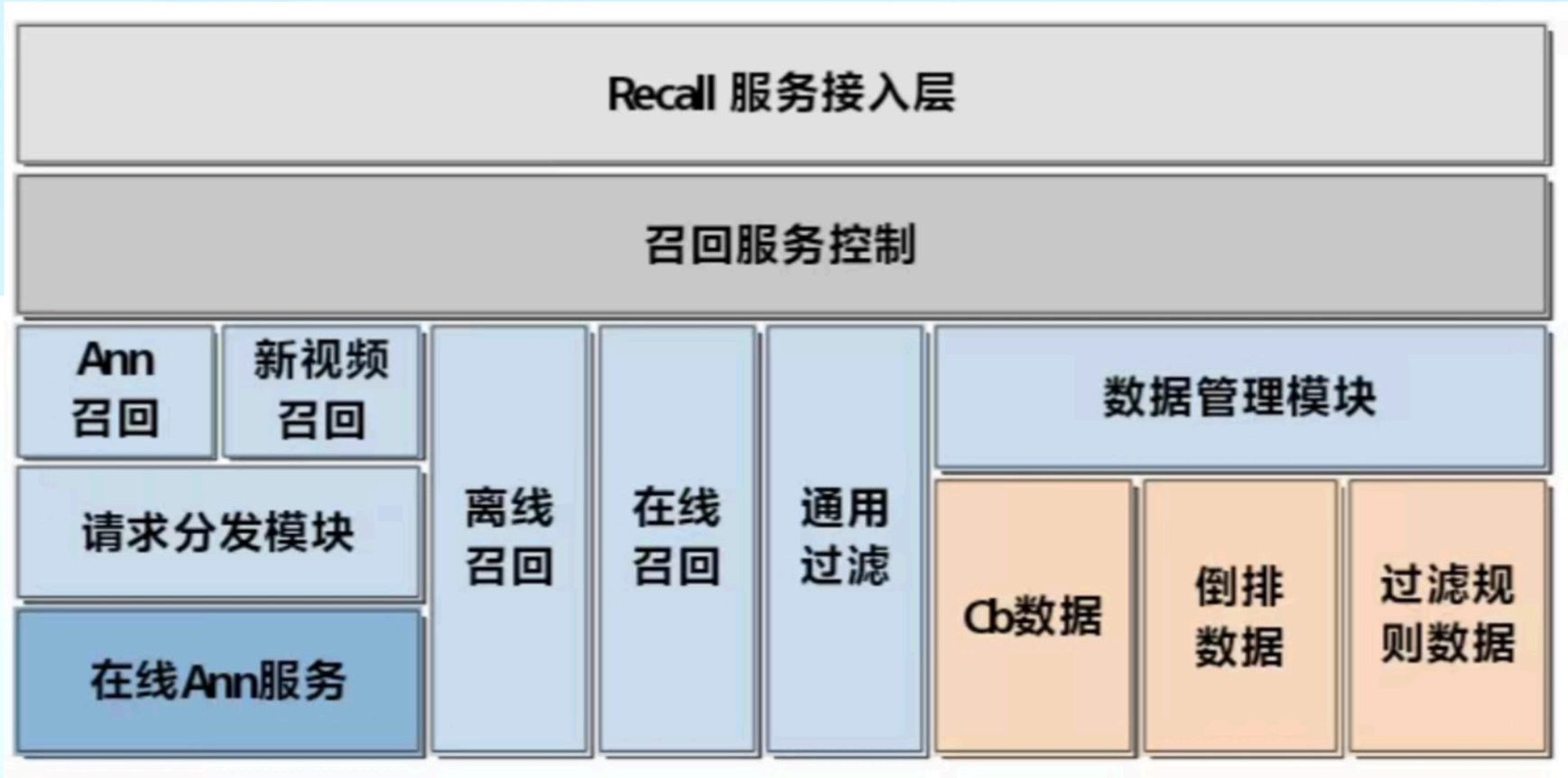
Module interpretation

corpus services

- The Corpus service is responsible for loading and maintaining corpus data, and providing services related to corpus data to the outside world. The Corpus service mainly provides two types of external interfaces: data query interface and filtering interface.

Module interpretation

recall service



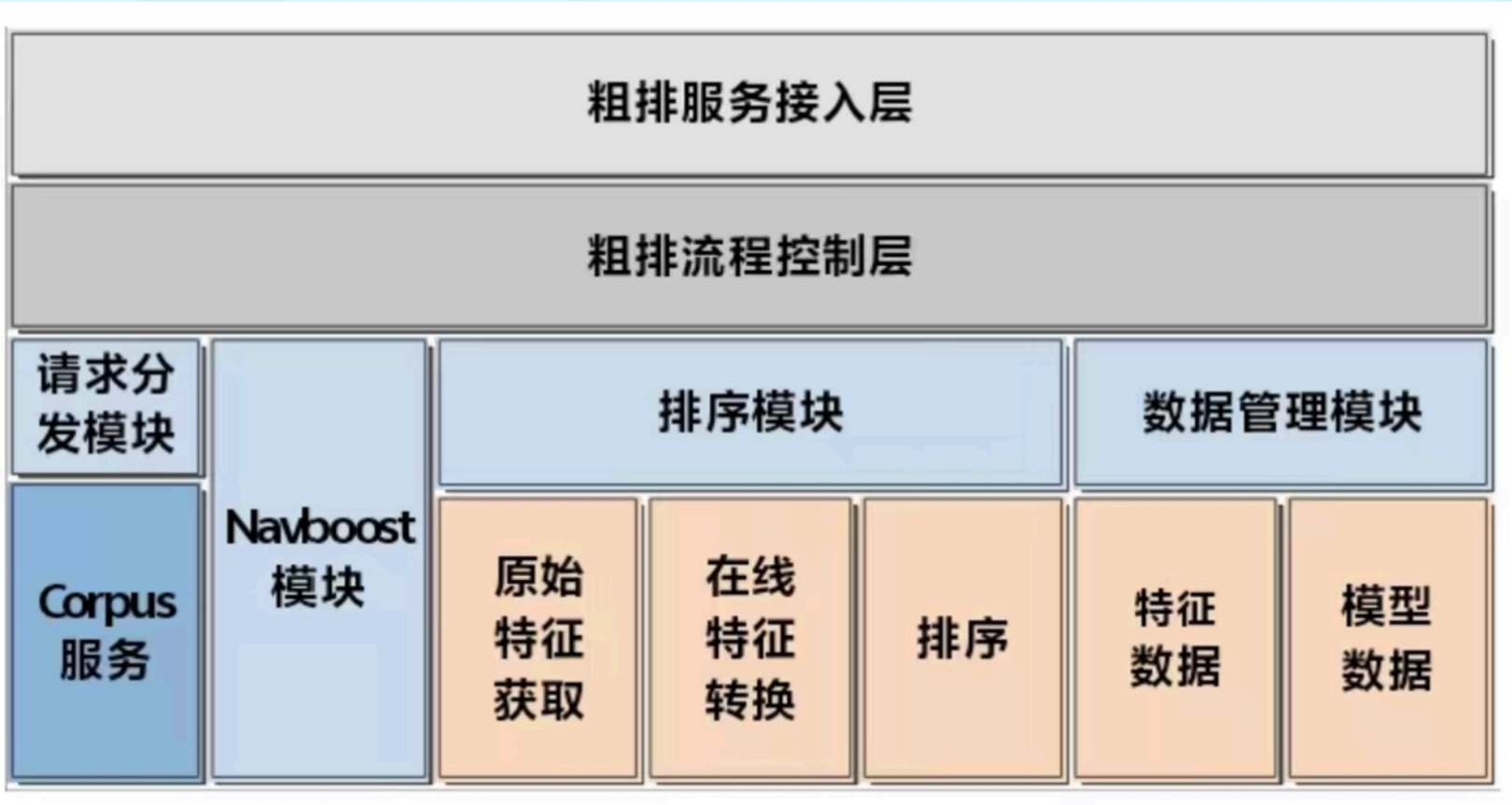
Module interpretation

recall service

- The recall module provides all recall services. The recall control information of the relevant engines is configured in the master's bucket configuration file. Different buckets support different recall configurations (which may be different recall sources, recall proportions, recall filter conditions, etc.). The recall configuration information is carried by the interface when the master accesses the recall service. The recall service uses these configuration information to determine which recalls are required for the current request, and how to filter the recall data.

Module interpretation

roughrank service



Module interpretation

roughrank service

- The rough sorting service provides a simple sorting service for related engines. This module will use a simple ranking model, or directly use some formula to rank the recall collection. The purpose of this module is to reduce the size of the candidate set, thereby reducing the size of the candidate set entering fine ranking, thereby reducing the calculation amount of the fine ranking service.

Module interpretation

- Navboost modules and principles

- Navboost module: A simple rough sorting method that uses the navboost formula to calculate scores to sort the recall set.

- NavBoost is an algorithm used in recommendation systems, especially in enhancing the performance of collaborative filtering recommendation systems. The principle of NavBoost (Navigation Boosting) is to improve the accuracy of recommendations and user satisfaction by combining user navigation behavior and traditional collaborative filtering methods.

Module interpretation

- Navboost modules and principles

- User navigation behavior: records the user's browsing path and behavior data on the website, including clicks, browsing time, page jumps, etc. This data can reveal users' interests and preferences.
- Collaborative filtering: Traditional collaborative filtering methods, including user-based and item-based collaborative filtering, use user historical behavior data (such as ratings, purchase records) to make recommendations.
- Weighted combination: NavBoost weighted combination of user navigation behavior and collaborative filtering results to generate the final recommendation results. By analyzing users' real-time behavioral data, recommendation results can be dynamically adjusted to make them more personalized and timely.

Module interpretation

- Navboost modules and principles

NavBoost 的公式

假设我们有两个主要的推荐来源：

- $CF(u, i)$：用户 u 对物品 i 的协同过滤推荐得分
- $NB(u, i)$：用户 u 对物品 i 的导航行为推荐得分

NavBoost 的推荐得分可以表示为这两个得分的加权和：

$$Score(u, i) = \alpha \cdot CF(u, i) + (1 - \alpha) \cdot NB(u, i)$$

- 其中，α 是一个权重参数，用于平衡协同过滤和导航行为的影响。

- Real-time: Combined with the user's real-time navigation behavior, the recommendation results can be dynamically adjusted to improve the real-time and relevance of the recommendations.
- Personalization: comprehensively consider the user's historical behavior and real-time behavior to provide more personalized recommendation results.
- Accuracy: By weighted combination of different recommendation sources, the accuracy of recommendations and user satisfaction can be improved.

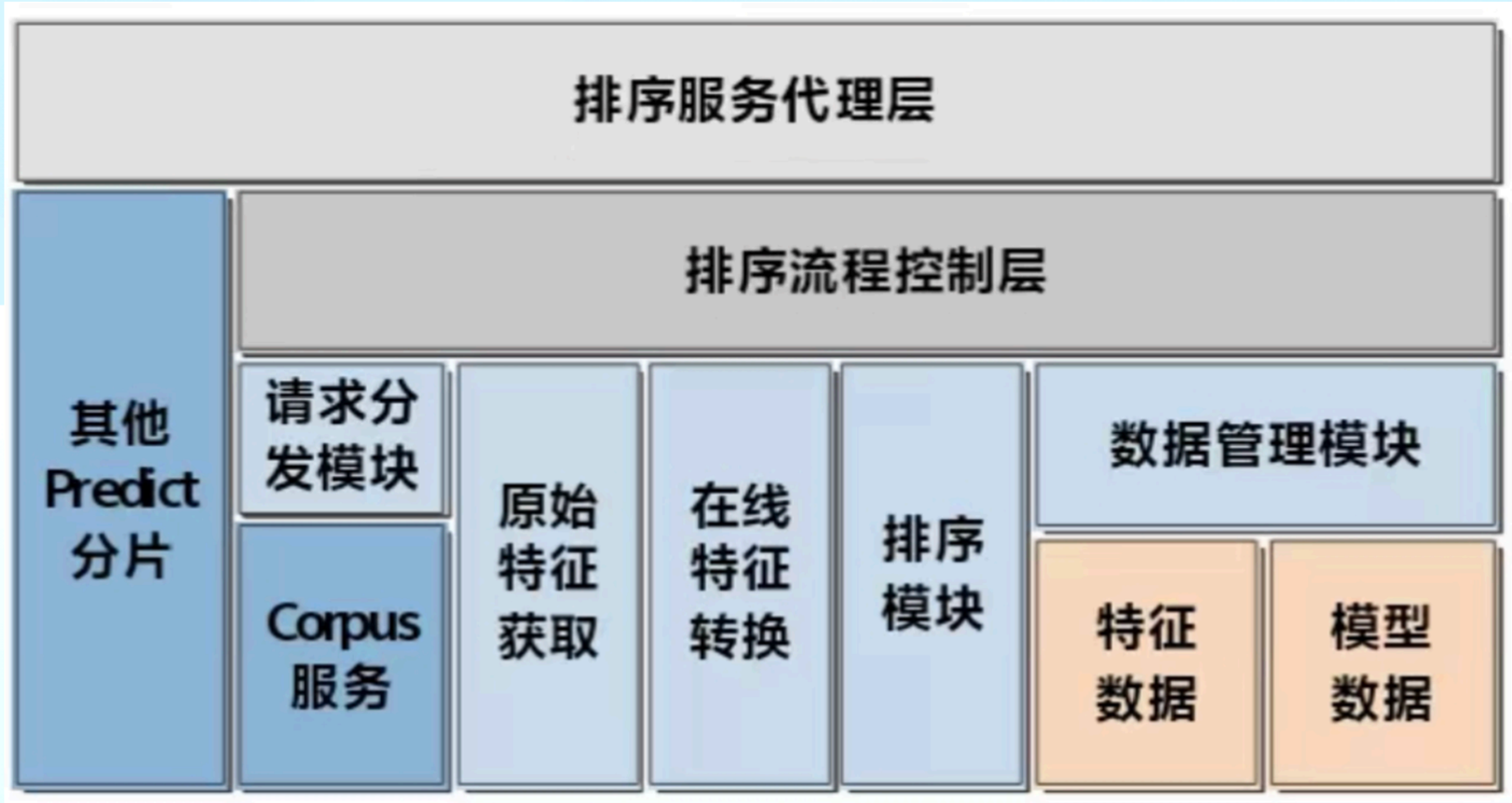
Module interpretation

Rough Ranking

- Sorting module: rough sorting method, using a simple sorting model (such as gbdt+lr), and constructing simple features to sort the recall set.

Module interpretation

predict service



Module interpretation

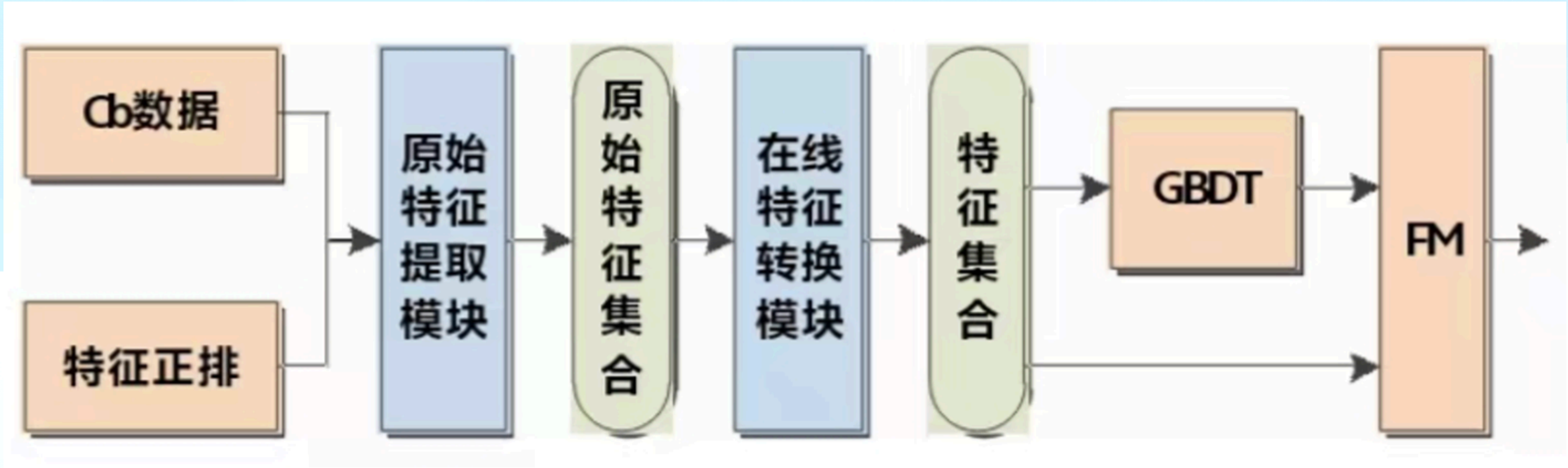
predict service

- The Predict service provides a more complex and accurate sorting service for related engines. Compared with the rough sorting service, this service uses more features and more complex models, so the sorting results will be more accurate than the rough sorting module, but it also consumes more computing resources. . The Predict module is called by the master service. The calling time is after the master receives the response from the rough sorting service and trims the result set returned by the rough sorting.

- The sorting module supports gbdt+lr, gbdt+fm, dnn+fm, deepfm and other deep modules

Module interpretation

predict service

- 

- First input the feature set of gbdt into gbdt. After obtaining the output of gbdt, add the output to the feature set of fm. Finally, input the total set into the fm model, and the fm will calculate the final result as the score.