



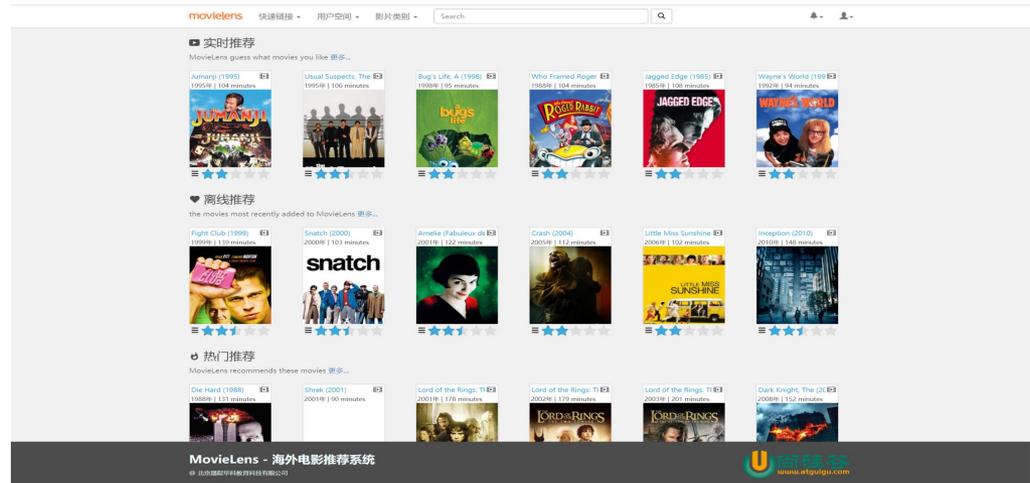
电影推荐系统设计

讲师：武晟然



主要内容

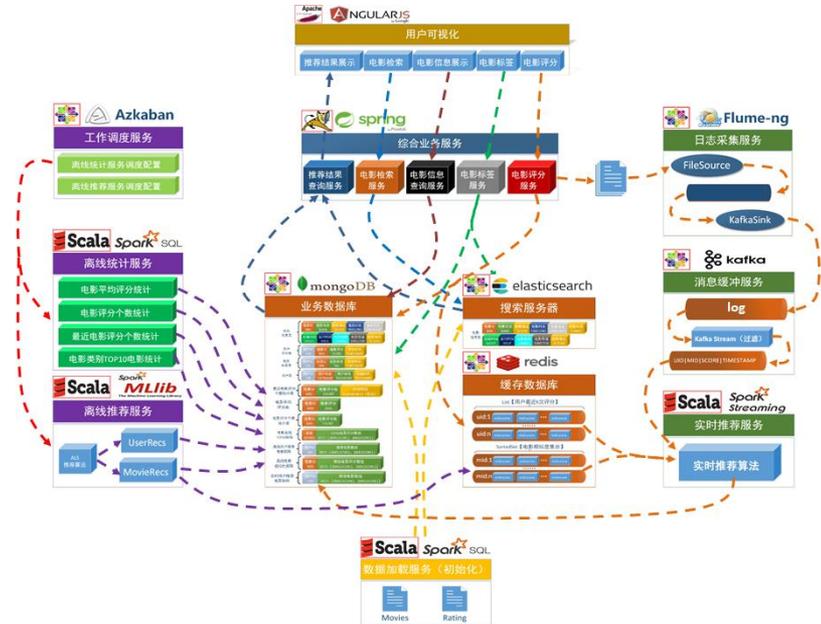
- 项目框架
- 数据源解析
- 统计推荐模块
- 离线推荐模块
- 实时推荐模块
- 基于内容的推荐模块





项目框架

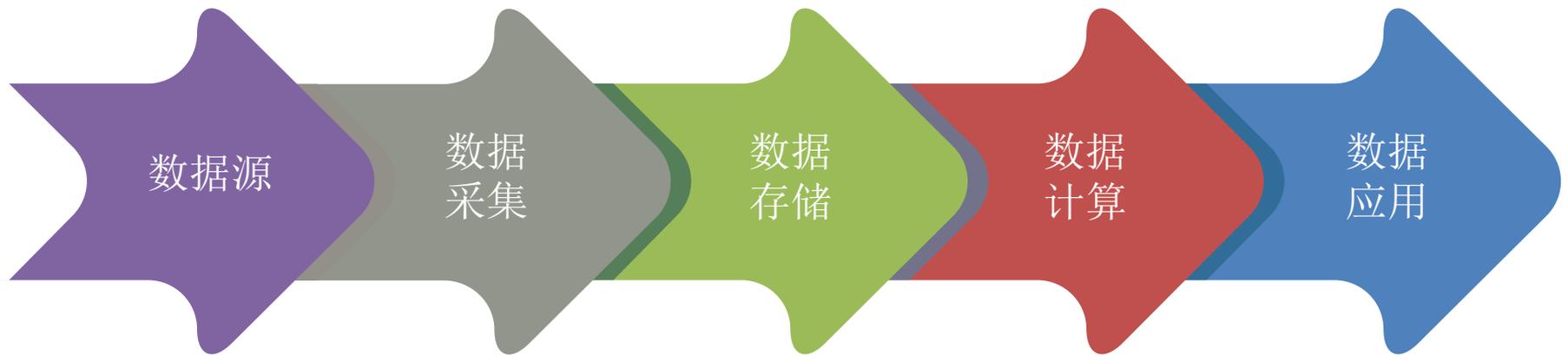
- 大数据处理流程
- 系统模块设计
- 项目系统架构
- 项目数据流图





数据生命周期

非结构化数据	图片视频	ETL工具	Oracle	Mahout	业务应用
		Scribe	GreenPlum	Storm	Tableau
半结构化数据	日志数据	Flume	Cassandra	Flink	BI分析
		Kafka	HBase	Spark	可视化 Echarts D3
结构化数据	关系数据	Sqoop	HDFS	MapReduce	





我们的目标

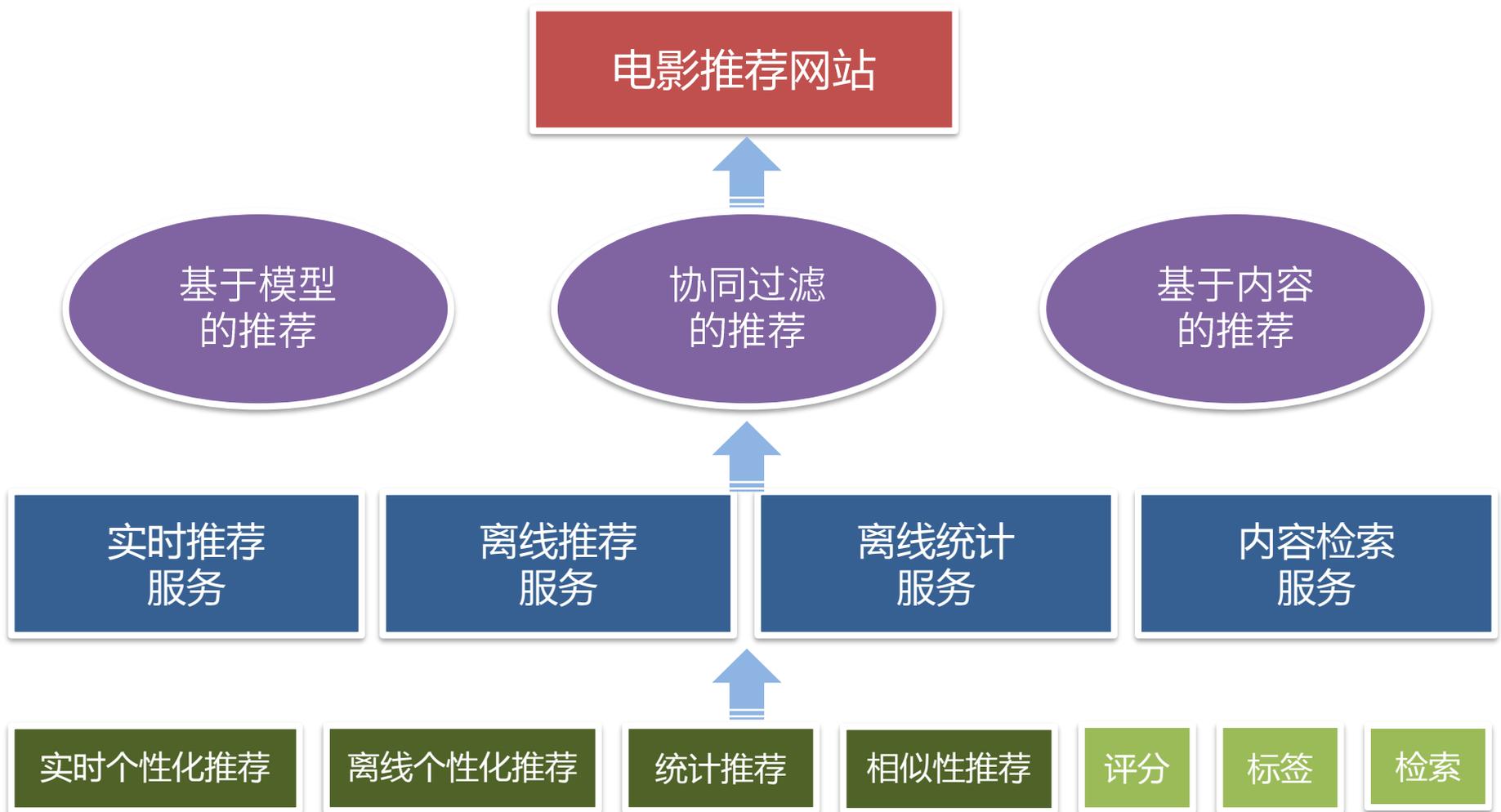
The screenshot displays the MovieLens website interface with several key sections highlighted by red circles:

- 实时推荐 (Real-time Recommendations):** Located at the top left, it features a header "movielens 快速链接" and "用户空间". Below it, the text reads "MovieLens guess what movies you like 更多...". It shows two movie cards: "Jumanji (1995)" and "Usual Suspects, Th".
- 离线推荐 (Offline Recommendations):** Below the real-time section, it says "the movies most recently added to MovieLens 更". It features two movie cards: "Fight Club (1999)" and "Snatch (2000)".
- 热门推荐 (Popular Recommendations):** At the bottom left, it says "MovieLens recommends these movies 更多...". It features two movie cards: "Die Hard (1988)" and "Shrek (2001)".
- Star Wars: Episode IV - A New Hope (1977) Detail Page:** The right side of the screenshot shows a detailed view of this movie. The title and year are circled in red. Below the title, there are star ratings, a plot summary: "Princess Leia is captured and held hostage by the evil Imperial forces in their effort to take over the galactic Empire. Venturesome Luke Skywalker and dashing captain Han Solo team together with the loveable robot duo R2-D2 and C-3PO to rescue the beautiful princess and restore peace and justice in the Empire.", and a list of cast and crew members.
- 我的标签 (My Tags):** A section for user tags with a search bar and a list of tags like "awesome soundtrack", "abc", "sci fi", "George Lucas", "stanam", "Science Fiction", "space", "critically acclaimed", "Sci-Fi", "action", "TV", "job", "space adventure", "coming of age", "science fiction", "classic", "sci fi", "abc".
- 相似推荐 (Similar Recommendations):** A section titled "MovieLens guess movies same with Star Wars: Episode IV - A New Hope (1977)". It shows a row of movie cards including "Monty Python and the Holy Grail", "Star Wars: Episode I - The Phantom Menace", "Raiders of the Lost Ark", "Star Wars: Episode II - Attack of the Clones", "Star Wars: Episode III - Revenge of the Sith", "Star Wars: Episode IV - A New Hope", "Star Wars: Episode V - The Empire Strikes Back", "Star Wars: Episode VI - Return of the Jedi", "The Untouchables", and "The Untouchables: The Motion Picture".

At the bottom of the page, there is a footer for "MovieLens - 海外电影推荐系统" and "尚硅谷 www.atguigu.com".

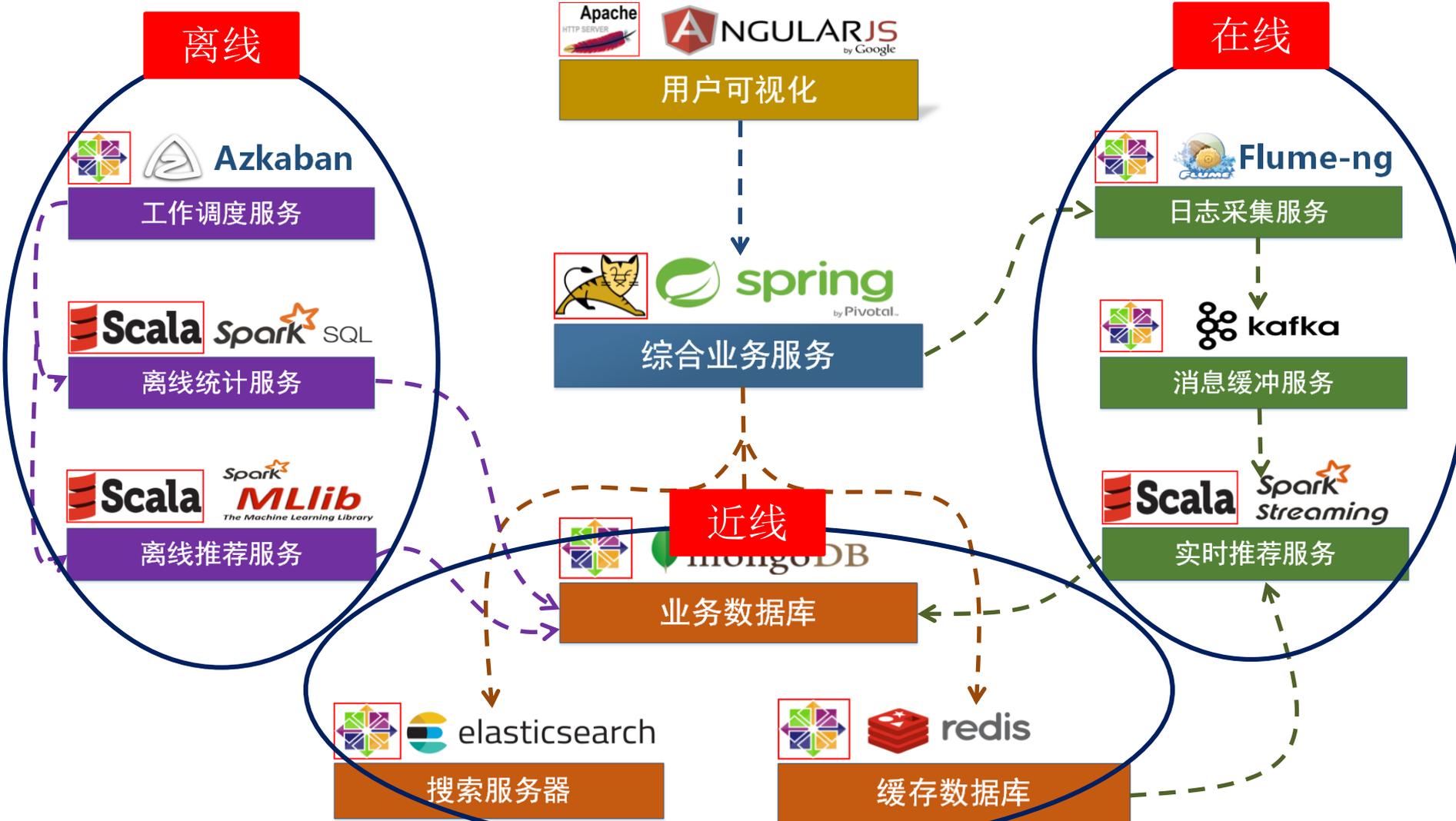


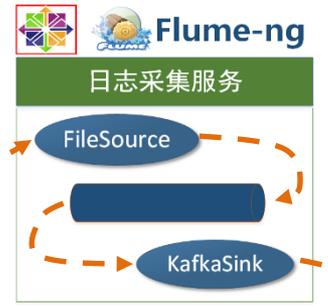
系统模块设计





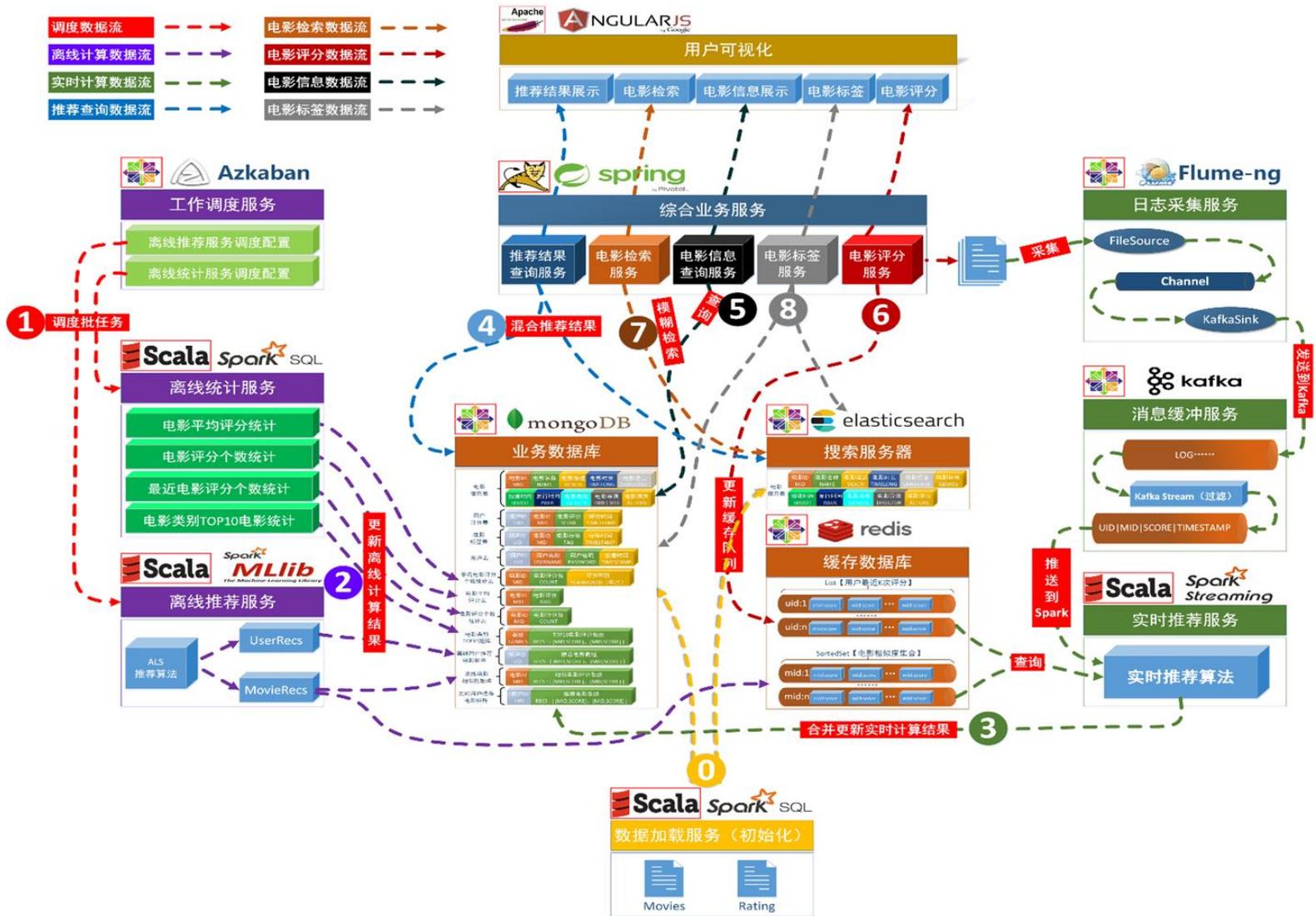
项目系统架构







系统数据流图





数据源解析

- 电影信息
- 用户评分信息
- 电影标签信息

movies.csv

ratings.csv

tags.csv



电影信息

电影 ID (MID)	电影名称 (NAME)	电影描述 (DESCRI)	电影时长 (TIMELO NG)	发行时间 (ISSUE)	拍摄时间 (SHOOT)	电影语言 (LANGUA GE)	电影类别 (DIRECT OR)	电影演员 (ACTOR S)	电影导演 (DIRECT OR)
1	Toy Story	-	81minutes	March 20 2001	1995	English	Adventure Animation ... Fantasy	Tom Hanks ... Jim	John Lasseter
...
30	Shanghai Triad	-	108minutes	December 12 2000	1995	Chinese	Crime Drama	Gong Li ... Li Bao-Tian	Zhang Yimou



用户评分信息

用户ID (UID)	电影ID (MID)	电影评分 (SCORE)	评分时间 (TIMESTAMP)
671	5816	4	1065111963
671	5902	3.5	1064245507
...
671	5952	5	1063502716

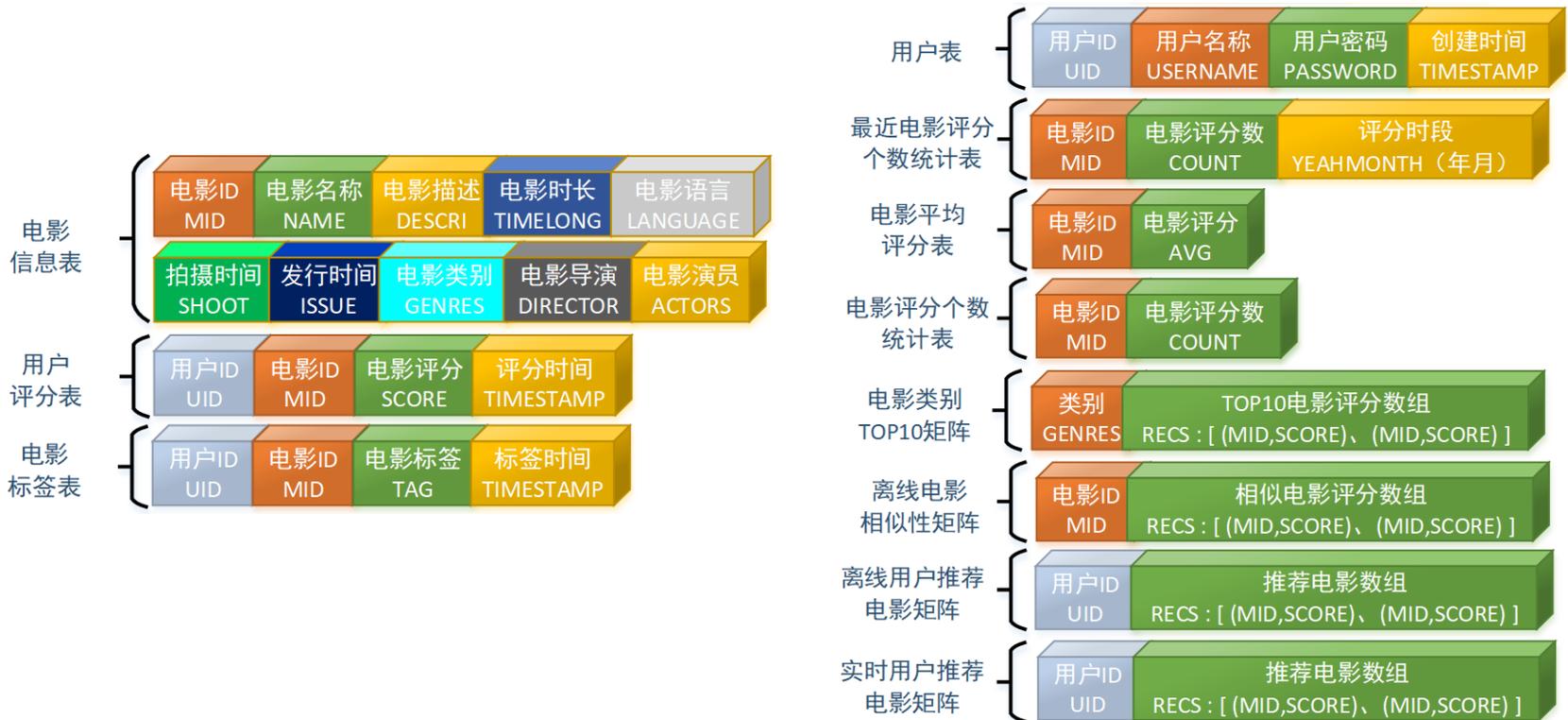


电影标签信息

用户ID (UID)	电影ID (MID)	电影标签 (TAG)	标签时间 (TIMESTAMP)
15	339	sandra 'boring' bullock	1138537770
15	1955	dentist	1193435061
...
15	100365	uganda	1425876220



主要数据模型



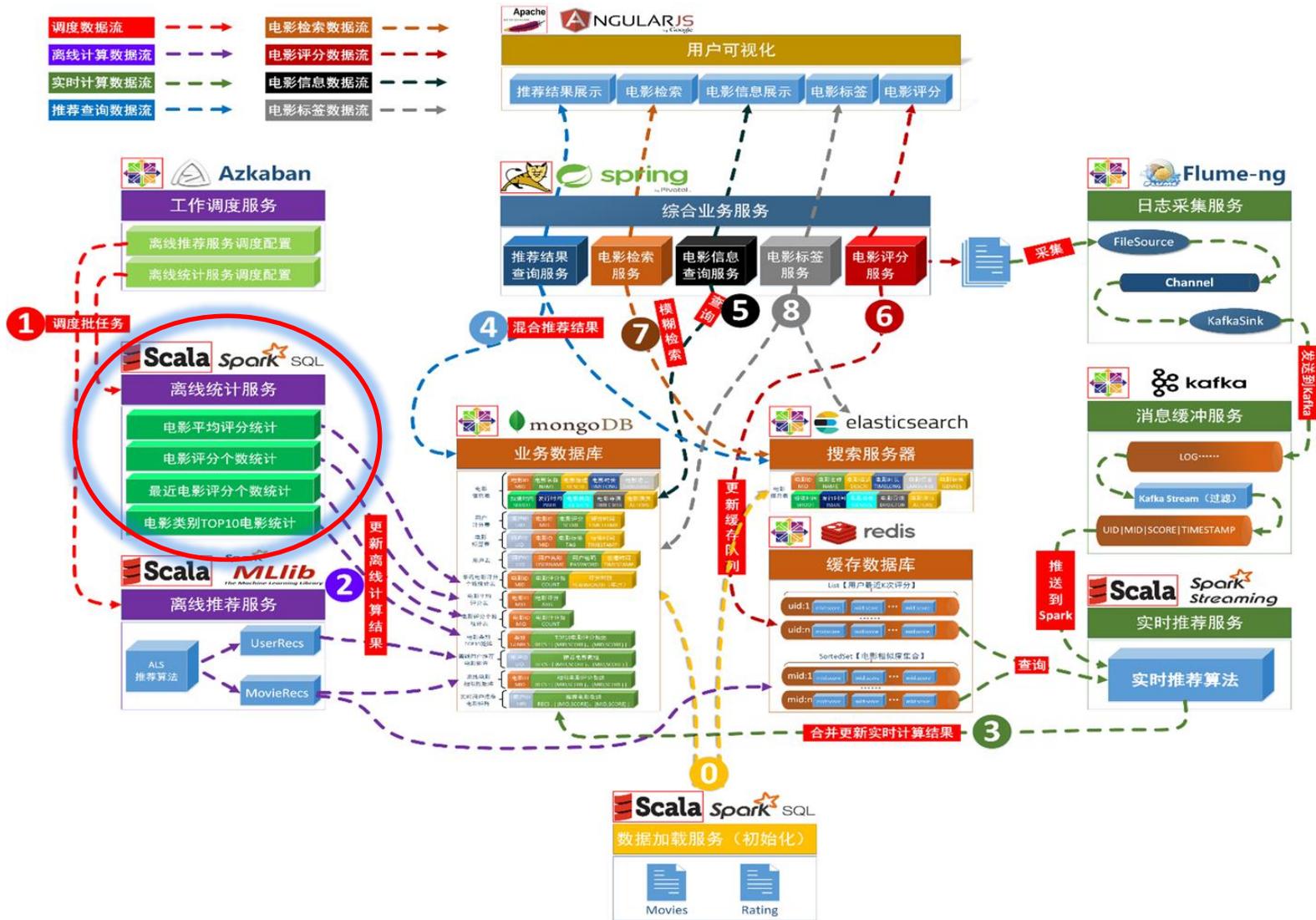


统计推荐模块

- 历史热门电影统计
- 近期热门电影统计
- 电影平均评分统计
- 各类别 Top10 优质电影统计

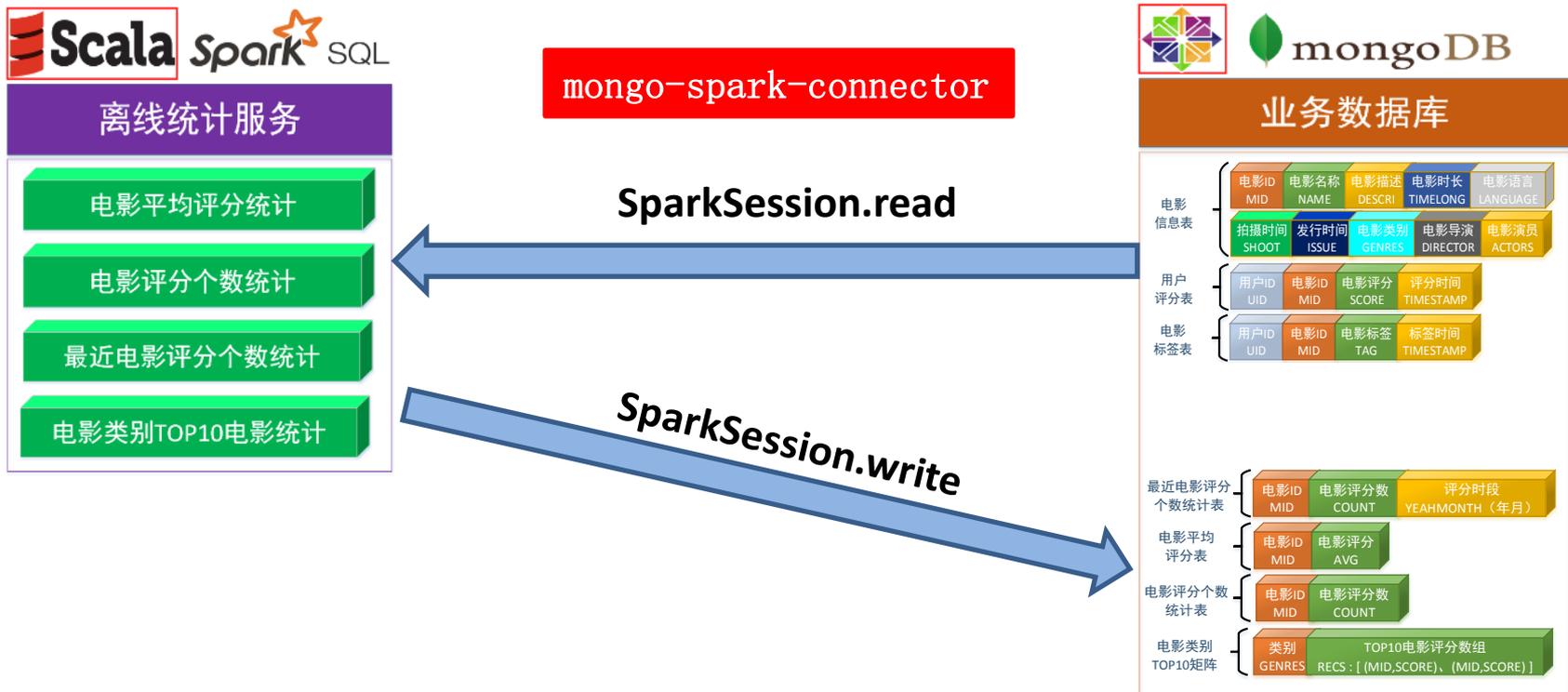


统计推荐模块





统计推荐模块





历史热门电影统计

- 统计所有历史数据中每个电影的评分数
- `select mid, count(mid) as count from ratings group by mid`

➡ RateMoreMovies

- RateMoreMovies 数据结构: mid, count



近期热门电影统计

- 统计每月的电影评分个数，就代表了电影近期的热门度
- `select mid, score, changeDate(timestamp) as yearmonth from ratings`
 ➔ `ratingOfMonth`
- `select mid, count(mid) as count ,yearmonth from ratingOfMonth group by yearmonth,mid order by yearmonth desc,count desc`
 ➔ **RateMoreRecentlyMovies**
- `changDate` : UDF函数，使用 `SimpleDateFormat` 对 `Date` 进行格式化，转化格式为 “yyyyMM”
- `RateMoreRecentlyMovies` 数据结构：mid, count, yearmonth



电影平均评分统计

- `select mid, avg(score) as avg from ratings group by mid`

➡ **AverageMovies**

- AverageMovies 数据结构: mid, avg



各类别 Top10 评分电影统计

- `select a.mid, genres, if(isnull(b.avg),0,b.avg) score from movies a left join averageMovies b on a.mid = b.mid`
➔ `movieWithScore`
- `spark.sql("select * from (select " +
"mid," +
"gen," +
"score, " +
"row_number() over(partition by gen order by score desc) rank " +
"from " +
"(select mid,score,explode(splitGe(genres)) gen from movieWithScore)
genresMovies) rankGenresMovies " +
"where rank <= 10")`
- `splitGe` : UDF函数, 按照 ‘|’ 字符对字符串进行切分

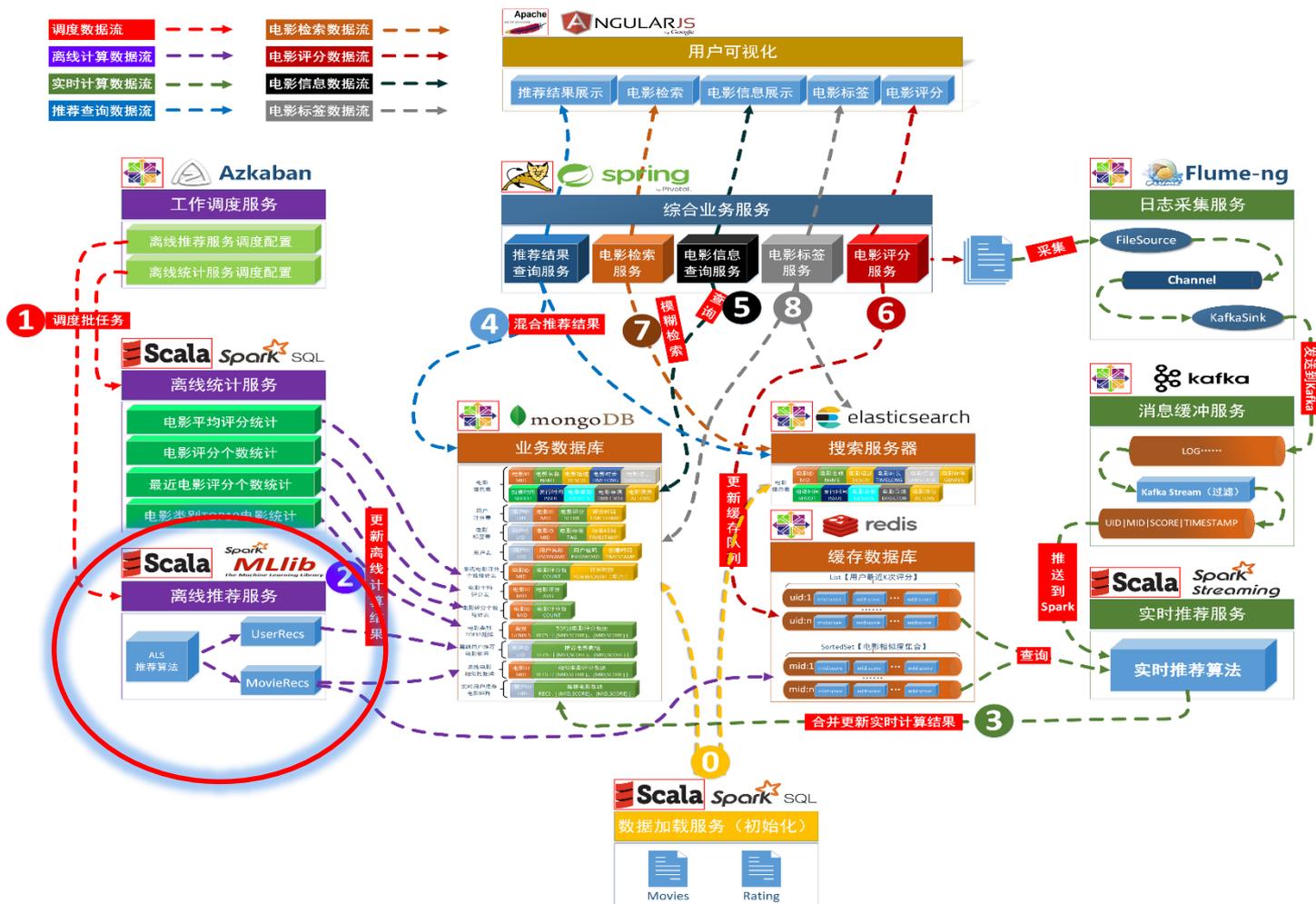


离线推荐模块

- 用ALS算法训练隐语义模型
- 计算用户推荐矩阵
- 计算电影相似度矩阵

离线推荐模块

电影推荐系统数据流图





ALS推荐模型训练



```
val model = ALS.train(trainData,rank,iterations,lambda)
```

- RMSE

均方根误差：均方误差的算术平方根，预测值与真实值之间的误差

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N (observed_t - predicted_t)^2}$$

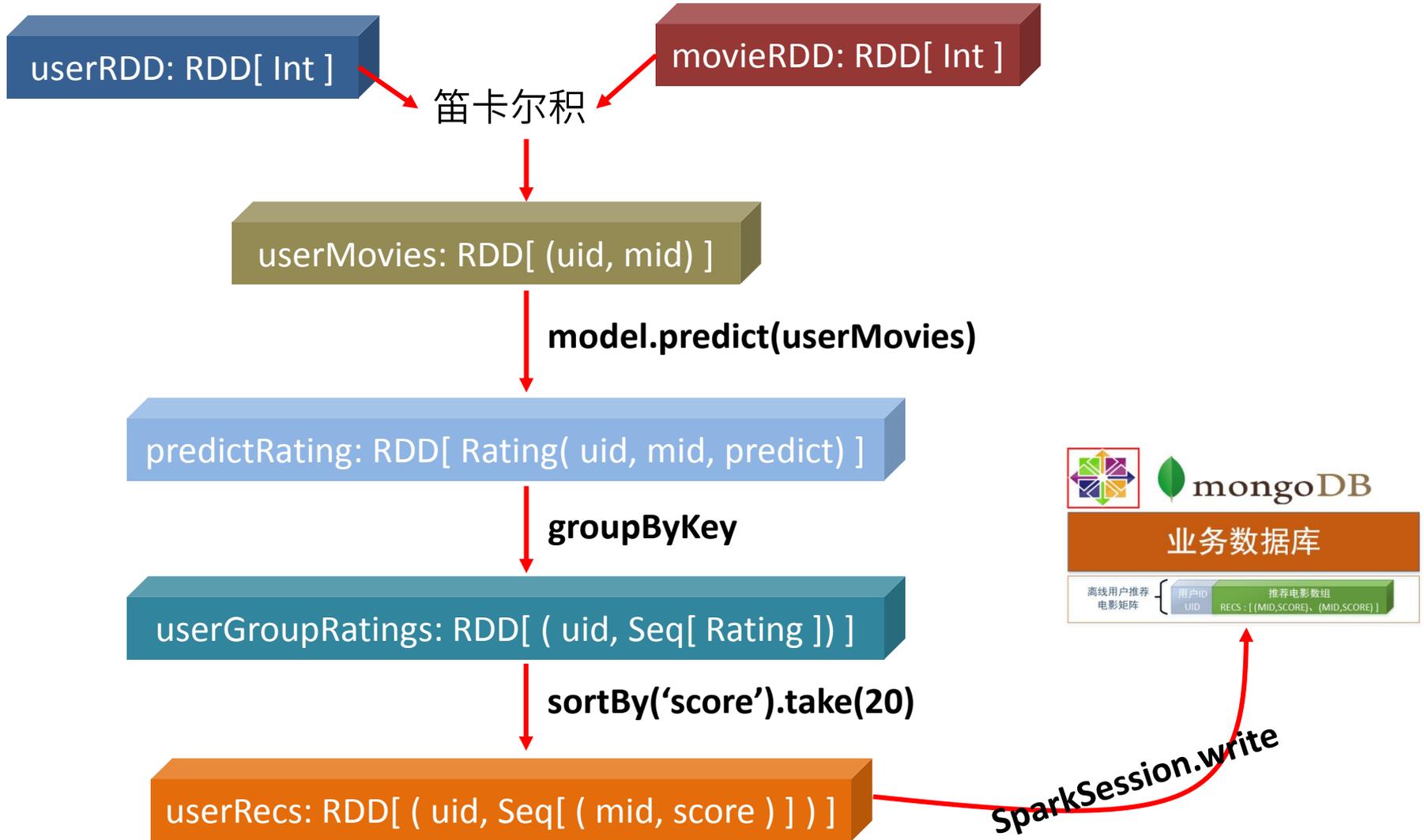
- 参数调整

可以通过均方根误差，来多次调整参数值，选择RMSE最小的一组参数值

- rank, iterations, lambda

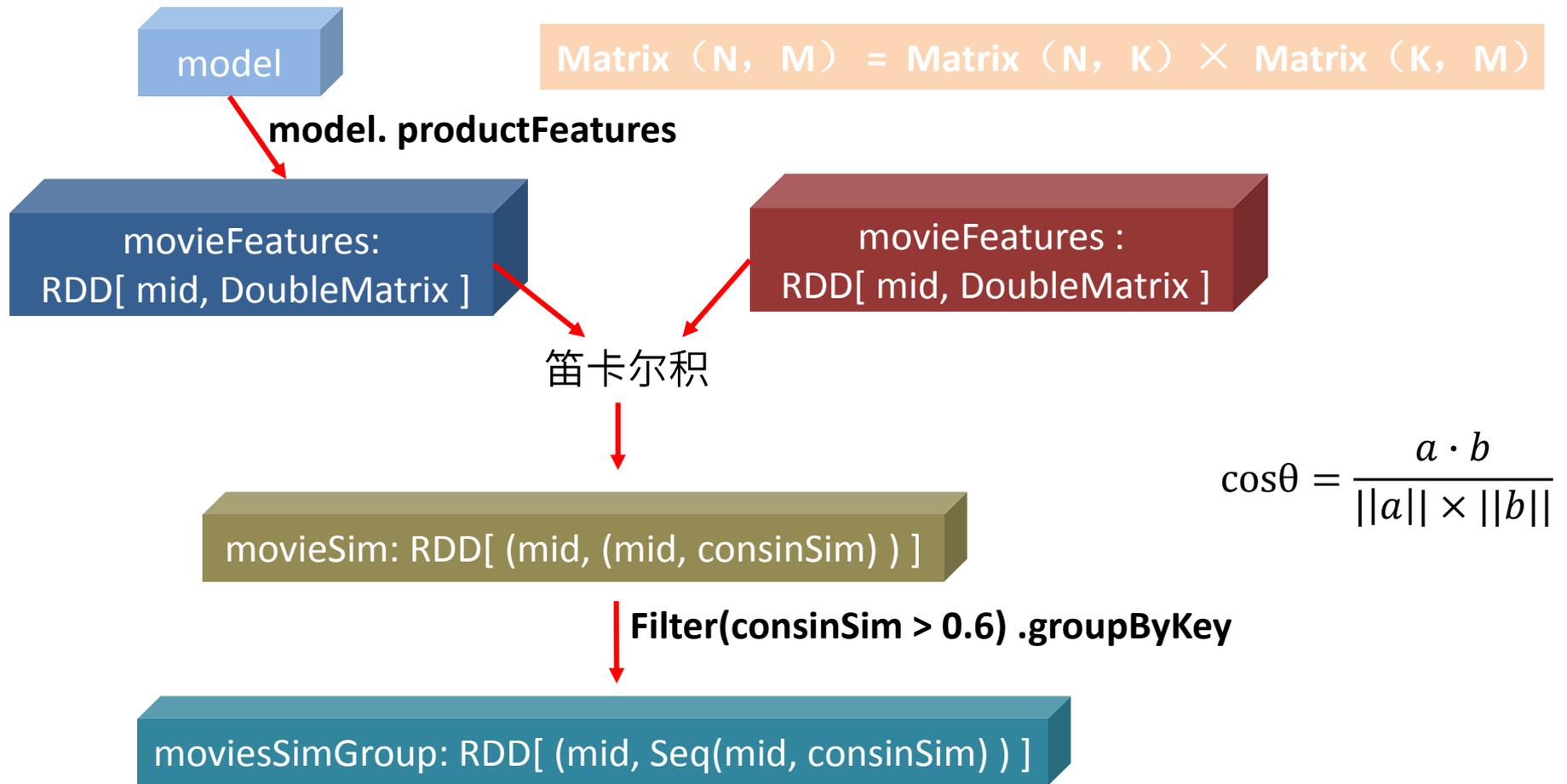


计算用户推荐矩阵



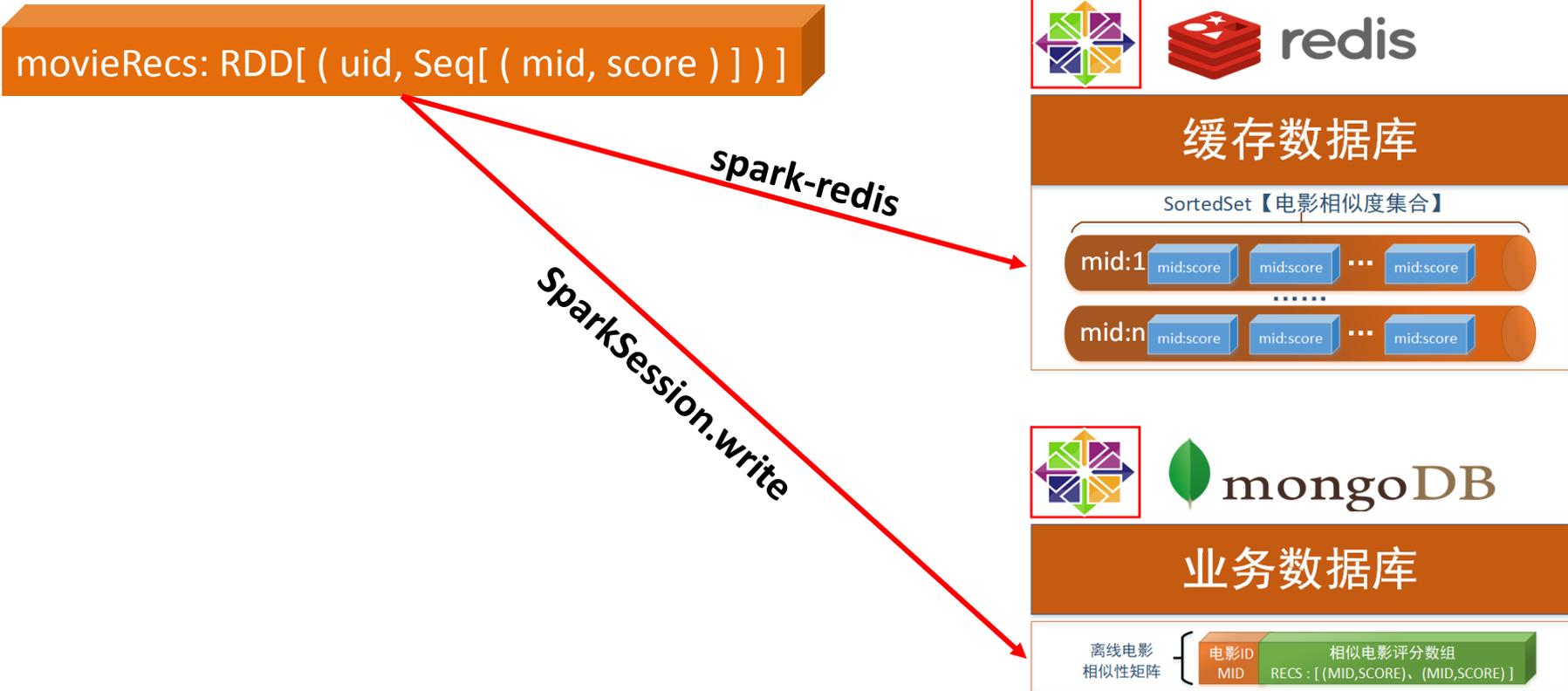


计算电影相似度矩阵





存储电影相似度矩阵



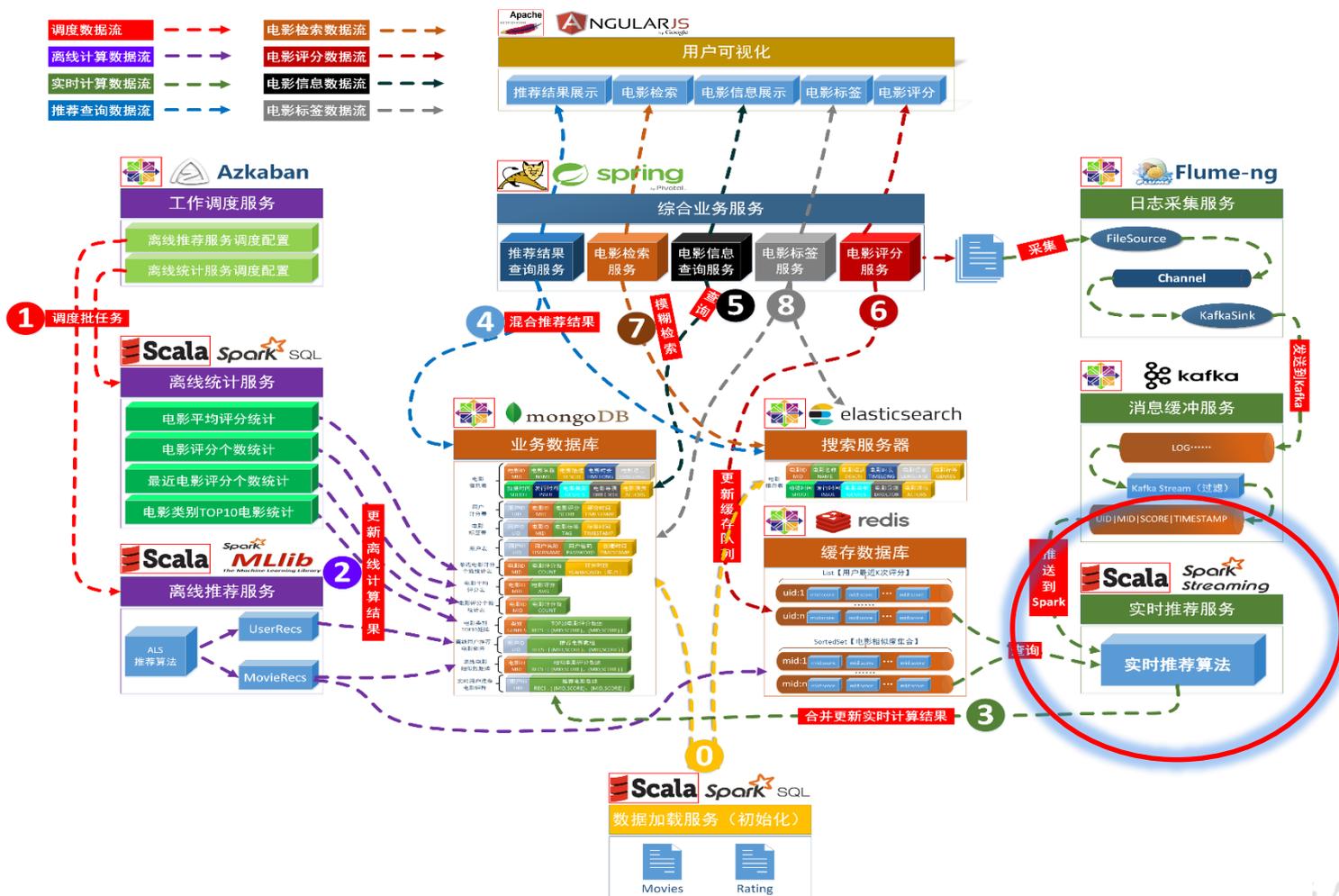


实时推荐模块

- 实时推荐架构
- 实时推荐优先级计算

基于模型的实时推荐模块

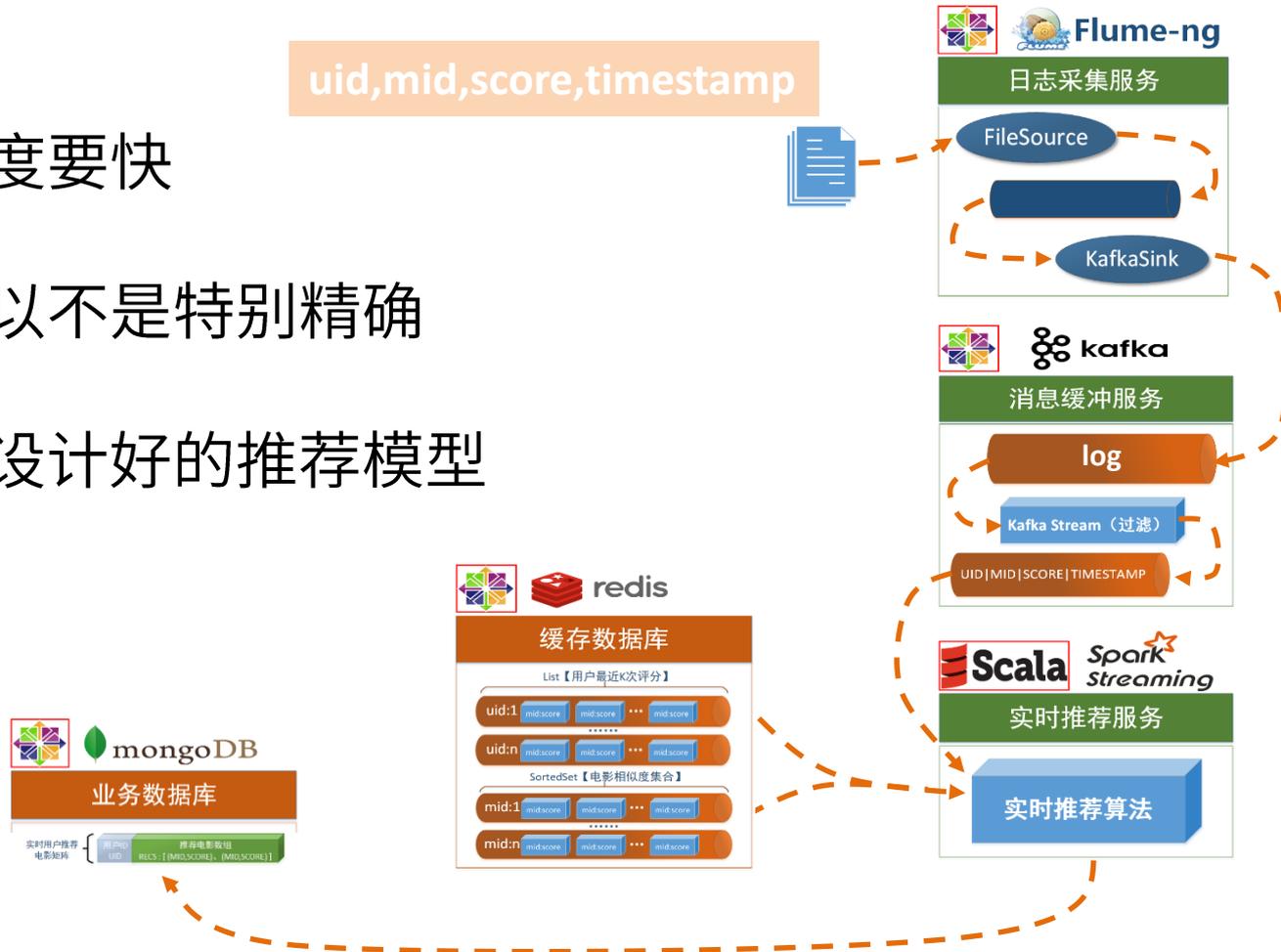
电影推荐系统数据流图





基于模型的实时推荐模块

- 计算速度要快
- 结果可以不是特别精确
- 有预先设计好的推荐模型





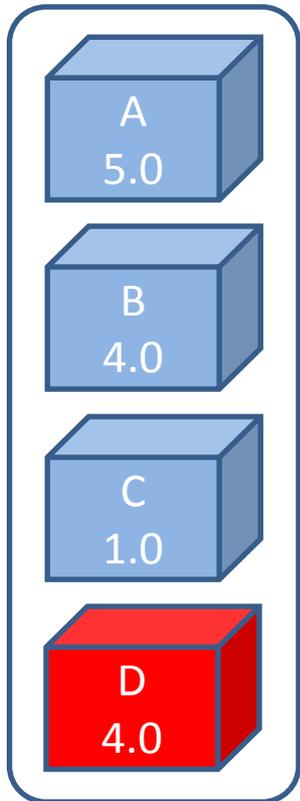
推荐优先级计算

基本原理：用户最近一段时间的口味是相似的

备选电影推荐优先级：

$$E_{uq} = \frac{\sum_{r \in RK} sim(q, r) \times R_r}{sim_sum} + \lg \max\{incount, 1\} - \lg \max\{recount, 1\}$$

用户最近 k 次评分



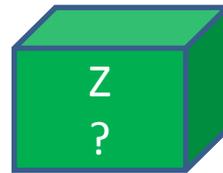
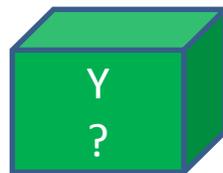
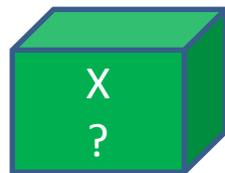
Sim(A, X)

Sim(B, X)

Sim(C, X)

X的推荐优先级分数为：

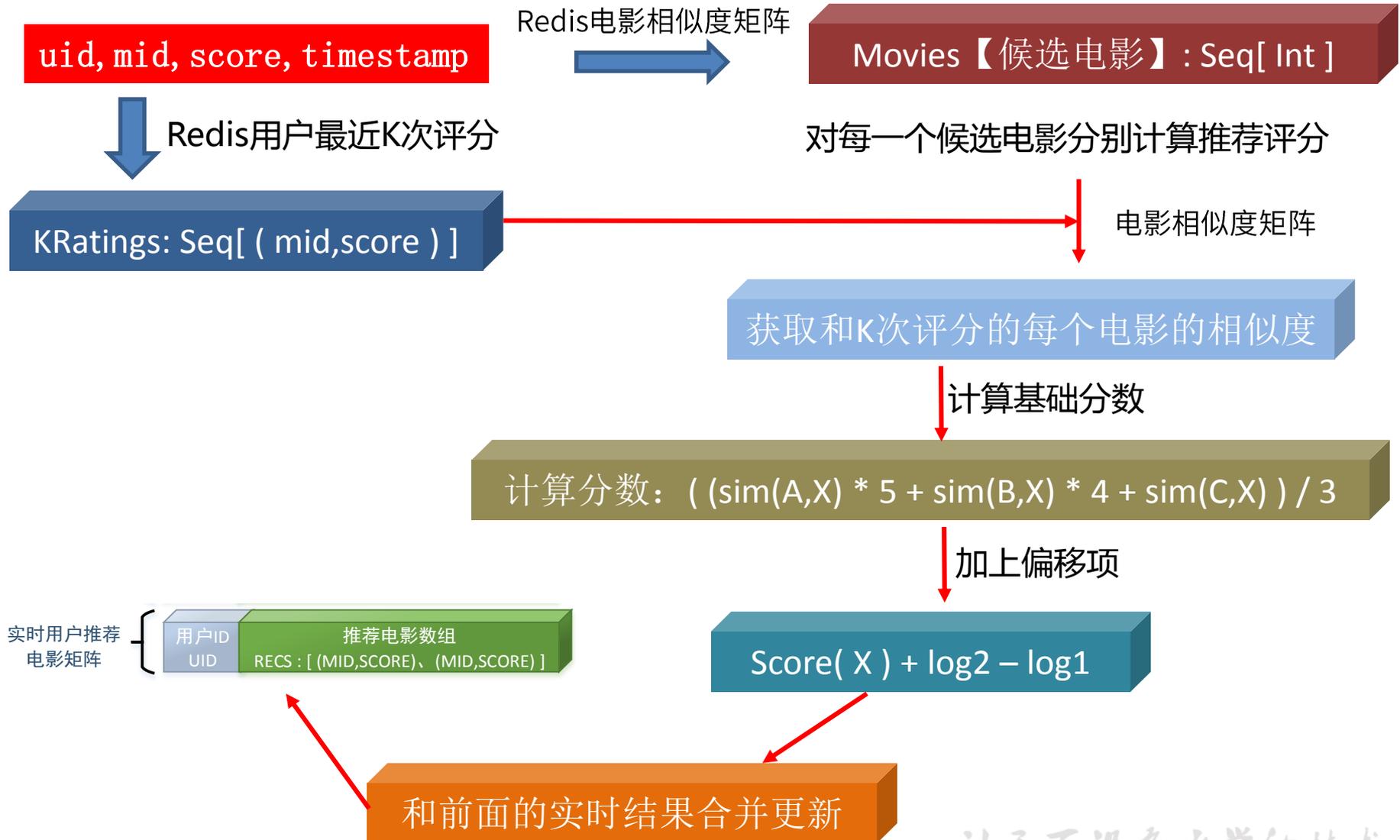
$$? = (sim(A, X) * 5 + sim(B, X) * 4 + sim(C, X) * 1) / 3 + \lg 2 - \lg 1$$



备选电影

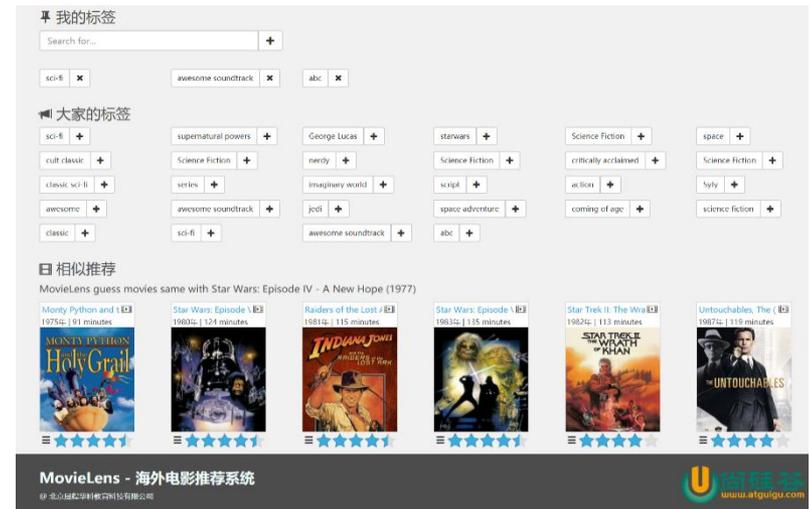


推荐优先级计算





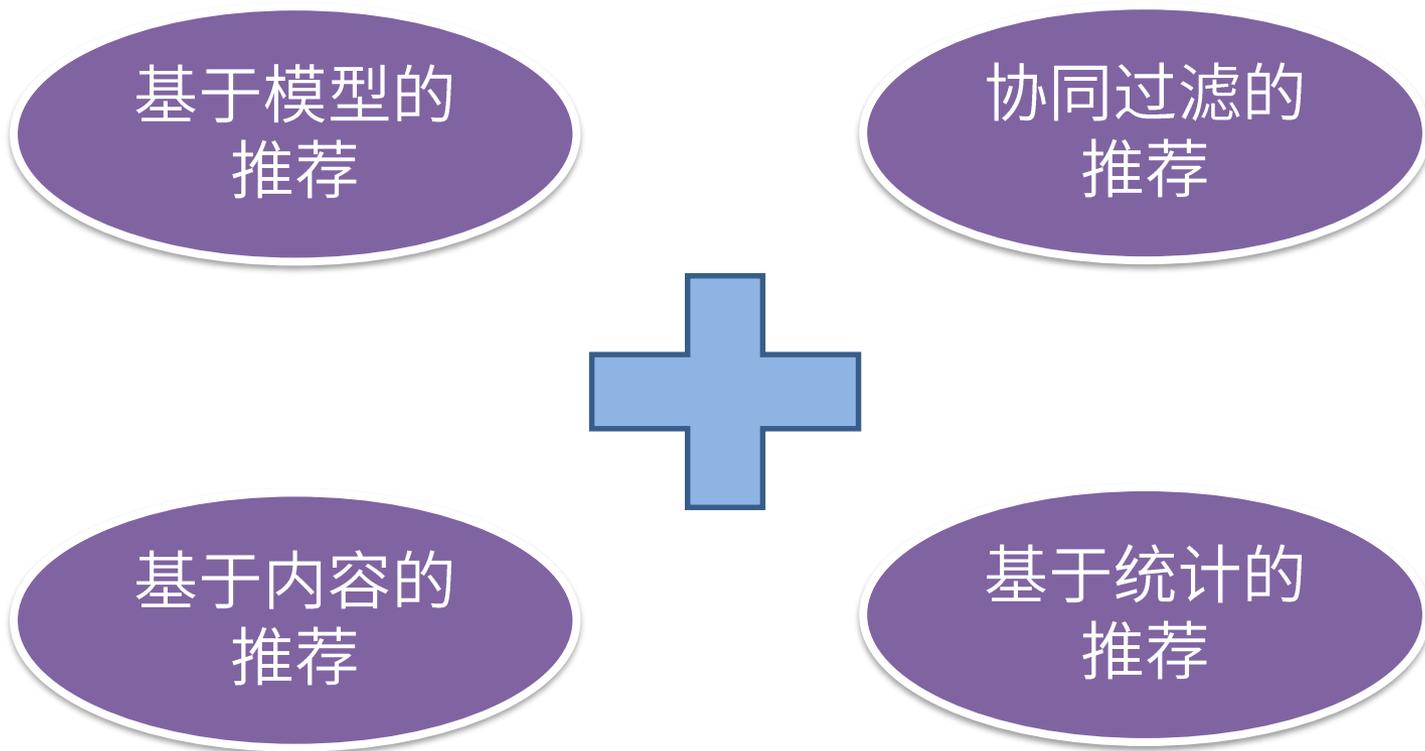
基于内容的推荐



- 电影 A 的相似电影？—— 有相同标签的电影
- Item-CF: 根据标签提取电影 A 的内容特征，选取与 A 特征相似的电影
- 根据 UGC 的特征提取 —— TF-IDF



混合推荐 —— 分区混合





Q & A