-- 大数据流式计算\_关键技术及系统实例\_孙大为.caj 完全可copy

-- 大数据系统和分析技术综述\_程学旗.caj

-- 大数据处理技术在智能交通中的应用\_周为钢.pdf

-- **Big Data Real-time Processing Based on Storm**

**-- Big-Data Analytics Challenges, Key Technologies and Prospects**

Generally the web mining tasks can be classified into threecategories: web content mining, web structure mining and webusage mining. In addition, there are two other differentapproaches to categorize web mining. In both, the abovecategories are reduced from three to two: web content miningand web usage mining. In one, web structure is treated as partof web Content while in the other; web usage is treated as partof web Structure.

--- Formation of Algorithms Module and Dynamic Interface in Web Mining and Storage Process to Retrieve Metadata

Web usage mining includes three main steps: Data

Preprocessing, Knowledge Extraction and analysis of

extracted results.

Data Log fetch

Data clean

Pageview Identification

Computing the Reference Length

User Identification

Session Identification

**A NewClustering and Preprocessing for Web Log Mining**

Semantic Data Mining

Web Usage Mining A Review on Process, Methods and Techniques --- 匹配度很高

RTIC-C: A Big Data System for Massive Traffic Information Mining – 较高

Toward Scalable Systems for Big Data Analytics:

A Technology Tutorial 最牛逼的文章

1. 可伸缩

在大数据流式计算环境中，系统的可伸缩性是制约大数据流式计算系统广泛应用的一个重要因素，一方面，流式数据的产生速率在高峰时期会不断增加且数据量巨大，持续时间往往很长，因此需要大数据流式系统具有很好的“可伸”的特征，可以实时适应数据增长的需求，实现对系统资源进行动态调整和快速部署，并保证整个系统的稳定性;另一方面，当流式数据的产生速率持续减少时，需要及时回收在高峰时期所分配的但目前已处于闲置或低效利用的资源，实现整个系统“可缩”的友好特征，并保障对用户是透明的。因此，系统中资源动态的配置、高效的组织、合理的布局、科学的架构和有效的分配，是保障整个系统可伸缩性的基础，同时，又尽可能地减少不必要的资源和能源的浪费。

a) 负载均衡

在大数据流式计算环境中，系统的负载均衡机制是制约系统稳定运行、高吞吐量计算、快速响应的一个关键因素。然而，当前多数系统不能有效地支持系统的负载均衡，如 Storm，S4 等系统均不支持负载均衡机制， Kafka系统实现了对负载均衡机制的部分支持:一方面，在大数据流式计算环境中，系统的数据速率具有明显的突变性，并且持续时间往往无法有效预测，这就导致在传统环境中具有很好的理论和实践效果的负载均衡策略在大数据流式计算环境中将不再适用;另一方面，当前大多数开源的大数据流式计算系统在架构的设计上尚未充分地、全面地考虑整个系统的负载均衡问题，在实践应用中，相关经验的积累又相对缺乏，因此，给大数据流式计算环境中负载均衡问题的研究带来了诸多实践中的困难和挑战。 大数据流式计算环境中的负载均衡问题的解决，需要结合具体的应用场景，系统地分析和总结隐藏在大数据流式计算中的数据流变化的基本特征和内在规律，结合传统系统负载均衡的经验，根据实践检验情况，不断进行相关机制的持续优化和逐步完善。

1. 容错