به نام خدا

محمد عذیری

220797072

1- بخش اول: هدوپ(Hadoop) چیست

هدوپ یک فریم‌ورک یا مجموعه‌ای از الگوریتم‌ها و لایبرری هاییست که با کمک آنها میتوان پردازش داده‌ای در سطح کلان انجام داد.

به طور کلی هدوپ از MR برای تحلیل داده‌ها استفاده می‌کند و این مفهوم با مدل مستر-اسلیو پیاده سازی شده است.

به طوری که یک job tracker داریم و تعدادی سرور به صورت task tracker به ما خدمت می‌کنند. این مقهوم با مدل ستاره‌ای پیاده شده است و کل کارها از سمت job tracker تعریف می‌شود. و سرور مستر به تمامی سرورهای دیگر متصل است و برای آنها وظایفشان را تعریف می‌کند.

این وظایف طبق مدل FIFO تعریف می‌شود و عملا تسک‌های تعریف شده در یک queue قرار می‌گیرند و برای سرور های تسک ترکر ارسال می‌شوند. و این بخش بخش توزیع یا Map بود.

در بخش بعدی سرورها داده‌های تحلیل شده را با دستور سرور مستر باهم می‌توانند ادغام کنند یا از داده‌های بین سروری استفاده کنند که تمامی این عملیات‌ها نیز تحت نظارت job tracker انجام می‌شود. و داده‌ها توکنایز شده یا به عبارتی map می‌شوند. حال مرحله map به اتمام رسیده و وارد فاز reduce می‌شویم.

ما داده‌ها را به صورت یک دیکشنری یا به عبارتی یک ساختار key-value داریم. حال می‌توانیم متناسب با کاری که مد نظر داریم از آنها بهره ببریم.

2- بخش دوم:اسپارک(spark) چیست

بر خلاف هدوپ که از روی دیسک ورودی را می‌گیرد، اسپارک ورودی را از حافظه اصلی یا مموری میخواند و سرعتی بسیار بالاتر از هدوپ دارد. همچنین می‌تواند برخلاف هدوپ در بخش MR، با تعداد سیستم‌های بسیار کمتری عملیات مورد نیاز را اجرا کند.

ولی چون مموری در دسترس غالبا از دیسک در دسترس کمتر است ممکن است در اندازه‌های بالا مارا دچار مشکل کند.

مقایسه اسپارک و هدوپ:

در بخش‌های مختلفی می‌توان این دو تکنولوژی را مقایسه کرد.

1 - سرعت:

- برای این بخش اسپارک با اختلاف از هدوپ برتری دارد (10 الی 100 برابر سریعتر) و مقدار سرعت هدوپ نیز رابطه مستقیم با سرعت هارد ما دارد.

2 - هزینه:

- بستگی دارد اگر صرفا این موضوع را بخواهیم در نظر بگیریم که ram از hard گرانتر است می‌توان ادعا کرد که هدوپ هزینه بسیار کمتری نسبت به اسپارک دارد ولی طبق صحبت‌های گفته شده می‌توان در تعداد سرورهای کمتری اسپارک را نسبت به هدوپ اجرا کرد پس بین هزینه در نظر گرفته شده برای حافظه و سرور می‌توان مثالی زد که هدوپ هزینه بیشتری داشته باشد.

3 - تحمل خطا:

- هدوپ یک سیستم بسیار مقاوم از نظر تحمل خطا است و می‌تواند بخاطر تعدد دیتاها در سراسر سرورهای متفاوت از آنها در صورت بروز مشکل استفاده کند و دیتایی میس نشود، همچنین در اسپارک هر مشکلی برای یک بلوک دیتا بوجود بیاید می‌تواند آن بلوک را بازسازی کند. و در اسپارک در صورت وجود خطا نیازی به راه اندازی مجدد برنامه نیست.

4 - مقدار دیتای مورد بررسی:

- برای بررسی جمعی دیتاها هدوپ بهترین گزینه است، بخاطر MR می‌توان از فایل‌های بزرگ برای بررسی استفاده کند ولی اسپارک محدودیت رم دارد.

5 - سهولت استفاده:

- برای کدنویسی بین اسپارک و هدوپ، هدوپ پیچیده‌تر و سخت‌تر است ولی اسپارک نسبتا یوزر فرندلی است.