**(1) Phục hồi dữ liệu**

Sao chép Hadoop cung cấp khả năng chịu lỗi tích hợp bảo vệ chống lại lỗi hệ thống. Ngay cả khi một nút bị lỗi, các ứng dụng vẫn có thể tiếp tục chạy trong khi tránh mất dữ liệu. HDFS lưu trữ các tệp trong các khối được sao chép để đảm bảo khả năng chịu lỗi, giúp cải thiện cả độ tin cậy và hiệu suất. Quản trị viên có thể định cấu hình kích thước khối và hệ số sao chép trên từng tệp.

**(2) Vị trí dữ liệu**

Hadoop di chuyển tính toán gần nơi dữ liệu lưu trú, thay vì di chuyển các tập dữ liệu lớn đến tính toán. Điều này giúp giảm tắc nghẽn mạng trong khi cải thiện thông lượng chung.

**(3) Có những bản phân phối Hadoop nào dành riêng cho từng nhà cung cấp?**

Một số nhà cung cấp hiện nay cung cấp các sản phẩm dựa trên Hadoop. Một số sản phẩm đáng chú ý hơn bao gồm:

Mây Trời.

Bản đồ R.

Amazon EMR (Bản đồ đàn hồi giảm).

Microsoft Azure HDInsight.

Máy chủ thông tin IBM InfoSphere.

Nền tảng dữ liệu Hortonworks.

**(4) Một số tệp cấu hình chính được sử dụng trong Hadoop là gì?**

Nền tảng Hadoop cung cấp nhiều tệp cấu hình để kiểm soát cài đặt cụm, bao gồm:

7adoop-env.sh. Biến môi trường dành riêng cho trang web để kiểm soát các tập lệnh Hadoop trong thư mục bin.

yarn-env.sh. Biến môi trường dành riêng cho trang web để kiểm soát các tập lệnh YARN trong thư mục bin.

mapred-site.xml. Thiết lập cấu hình dành riêng cho MapReduce, chẳng hạn như thiết lập MapReduce.framework.name.

core-site.xml. Thiết lập cấu hình cốt lõi, chẳng hạn như cấu hình I/O chung cho HDFS và MapReduce.

yarn-site.xml. Thiết lập cấu hình cụ thể cho ResourceManager và NodeManager của YARN.

hdfs-site.xml. Thiết lập cấu hình dành riêng cho HDFS, chẳng hạn như đường dẫn tệp nơi NameNode lưu trữ không gian tên và nhật ký giao dịch.

Hình: Biết các thành phần cốt lõi của Hadoop khi tham gia phỏng vấn dữ liệu lớn.

Tìm hiểu các thành phần cốt lõi của Hadoop -- HDFS, YARN, MapReduce và Hadoop Common -- khi chuẩn bị cho buổi phỏng vấn về dữ liệu lớn.

**(5) HDFS là gì và các thành phần chính của nó là gì?**

HDFS là một hệ thống tệp phân tán đóng vai trò là môi trường lưu trữ mặc định của Hadoop. Nó có thể chạy trên phần cứng hàng hóa giá rẻ, đồng thời cung cấp mức độ chịu lỗi cao. HDFS lưu trữ nhiều loại dữ liệu khác nhau trong một môi trường phân tán cung cấp thông lượng cao cho các ứng dụng có bộ dữ liệu lớn. HDFS được triển khai trong kiến ​​trúc chính/phụ, với mỗi cụm hỗ trợ hai loại nút chính sau:

NameNode. Một nút chính duy nhất quản lý không gian tên hệ thống tệp, điều chỉnh quyền truy cập của máy khách vào tệp và xử lý thông tin siêu dữ liệu cho tất cả các khối dữ liệu trong HDFS.

DataNode. Một nút phụ quản lý bộ lưu trữ được đính kèm vào mỗi nút trong cụm. Một cụm thường chứa nhiều phiên bản DataNode, nhưng thường chỉ có một DataNode cho mỗi nút vật lý. Mỗi DataNode phục vụ các yêu cầu đọc và ghi từ các máy khách của hệ thống tệp.

**(6) Hadoop YARN là gì và các thành phần chính của nó là gì?**

Hadoop YARN quản lý tài nguyên và cung cấp môi trường thực thi cho các quy trình cần thiết, đồng thời phân bổ tài nguyên hệ thống cho các ứng dụng đang chạy trong cụm. Nó cũng xử lý việc lập lịch và giám sát công việc. YARN tách biệt việc quản lý tài nguyên và lập lịch khỏi thành phần xử lý dữ liệu trong MapReduce.

YARN tách quản lý tài nguyên và lập lịch công việc thành hai daemon sau :

ResourceManager. Daemon này phân xử tài nguyên cho các ứng dụng của cụm. Nó bao gồm hai thành phần chính: Scheduler và ApplicationsManager. Scheduler phân bổ tài nguyên cho các ứng dụng đang chạy. ApplicationsManager có nhiều vai trò: chấp nhận các yêu cầu gửi công việc, đàm phán thực thi ApplicationMaster dành riêng cho ứng dụng và cung cấp dịch vụ để khởi động lại vùng chứa ApplicationMaster khi xảy ra lỗi.

NodeManager. Daemon này khởi chạy và quản lý các container trên một nút và sử dụng chúng để chạy các tác vụ được chỉ định. NodeManager cũng chạy các dịch vụ xác định tình trạng của nút, chẳng hạn như thực hiện kiểm tra đĩa. Hơn nữa, NodeManager có thể thực hiện các tác vụ do người dùng chỉ định.

**(7) Các chế độ hoạt động chính của Hadoop là gì?**

Hadoop hỗ trợ ba nút hoạt động chính.

Standalone. Còn được gọi là chế độ Local, chế độ Standalone là chế độ mặc định. Nó chạy như một tiến trình Java duy nhất trên một nút duy nhất. Nó cũng sử dụng hệ thống tệp cục bộ và không yêu cầu thay đổi cấu hình. Chế độ Standalone chủ yếu được sử dụng cho mục đích gỡ lỗi.

Phân phối giả. Cũng được gọi là cụm nút đơn, chế độ Phân phối giả chạy trên một máy duy nhất, nhưng mỗi daemon Hadoop chạy trong một quy trình Java riêng biệt. Chế độ này cũng sử dụng HDFS, thay vì hệ thống tệp cục bộ và yêu cầu thay đổi cấu hình. Chế độ này thường được sử dụng cho mục đích gỡ lỗi và thử nghiệm.

Phân phối hoàn toàn. Đây là chế độ sản xuất đầy đủ, với tất cả các daemon chạy trên các nút riêng biệt trong cấu hình chính/phụ. Dữ liệu được phân phối trên toàn bộ cụm, có thể dao động từ một vài nút đến hàng nghìn nút. Chế độ này yêu cầu thay đổi cấu hình nhưng cung cấp khả năng mở rộng, độ tin cậy và khả năng chịu lỗi như mong đợi của một hệ thống sản xuất.

**(8) Ba định dạng đầu vào phổ biến trong Hadoop là gì?**

Hadoop hỗ trợ nhiều định dạng đầu vào, xác định hình dạng của dữ liệu khi dữ liệu được thu thập vào nền tảng Hadoop. Các định dạng đầu vào sau đây là ba trong số những định dạng phổ biến nhất:

Văn bản. Đây là định dạng đầu vào mặc định. Mỗi dòng trong tệp được coi là một bản ghi riêng biệt. Các bản ghi được lưu dưới dạng cặp khóa/giá trị, với dòng văn bản được coi là giá trị.

Văn bản khóa-giá trị. Định dạng đầu vào này tương tự như định dạng văn bản, chia mỗi dòng thành các bản ghi riêng biệt. Không giống như định dạng văn bản, coi toàn bộ dòng là giá trị, định dạng văn bản khóa-giá trị chia chính dòng thành khóa và giá trị, sử dụng ký tự tab làm dấu phân cách.

Tệp trình tự . Định dạng này đọc các tệp nhị phân lưu trữ trình tự các cặp khóa-giá trị do người dùng xác định dưới dạng các bản ghi riêng lẻ.

Hadoop cũng hỗ trợ các định dạng đầu vào khác, vì vậy hãy hiểu rõ về chúng.

**(9) Điều gì làm cho môi trường HDFS có khả năng chịu lỗi?**

HDFS dễ dàng sao chép dữ liệu sang các DataNode khác nhau. HDFS chia nhỏ các tệp thành các khối được phân phối trên các nút trong cụm. Mỗi khối cũng được sao chép sang các nút khác. Nếu một nút bị lỗi, các nút khác sẽ tiếp quản, cho phép các ứng dụng truy cập dữ liệu thông qua một trong các nút sao lưu.

**(10) Nhận thức về giá đỡ trong cụm Hadoop là gì?**

Nhận thức về giá là một trong những cơ chế được Hadoop sử dụng để tối ưu hóa quyền truy cập dữ liệu khi xử lý các yêu cầu đọc và ghi của máy khách. Khi có yêu cầu, NameNode sẽ xác định và chọn DataNode gần nhất, tốt nhất là những DataNode trên cùng giá hoặc trên các giá gần đó. Nhận thức về giá có thể giúp cải thiện hiệu suất và độ tin cậy, đồng thời giảm lưu lượng mạng. Nhận thức về giá cũng có thể đóng vai trò trong khả năng chịu lỗi. Ví dụ, NameNode đặt các bản sao khối dữ liệu trên các giá riêng biệt để đảm bảo quyền truy cập liên tục trong trường hợp bộ chuyển mạch mạng bị lỗi hoặc giá không khả dụng vì những lý do khác.

**(11) Hadoop bảo vệ dữ liệu khỏi truy cập trái phép như thế nào?**

Hadoop sử dụng giao thức xác thực mạng Kerberos để bảo vệ dữ liệu khỏi truy cập trái phép. Kerberos sử dụng mật mã khóa bí mật để cung cấp xác thực mạnh cho các ứng dụng máy khách/máy chủ. Máy khách phải trải qua ba bước cơ bản sau -- mỗi bước bao gồm trao đổi tin nhắn với máy chủ -- để chứng minh danh tính của mình với máy chủ đó:

Xác thực. Máy khách gửi yêu cầu xác thực đến máy chủ xác thực Kerberos. Máy chủ xác minh máy khách và gửi cho máy khách một vé cấp vé (TGT) và một khóa phiên.

Ủy quyền. Sau khi xác thực, máy khách yêu cầu một phiếu dịch vụ từ máy chủ cấp phiếu (TGS). TGT phải được bao gồm trong yêu cầu. Nếu TGS xác thực máy khách, nó sẽ gửi phiếu dịch vụ và thông tin xác thực cần thiết để truy cập tài nguyên được yêu cầu.

Yêu cầu dịch vụ. Máy khách gửi yêu cầu của mình đến tài nguyên Hadoop mà nó đang cố gắng truy cập. Yêu cầu phải bao gồm phiếu dịch vụ do TGS phát hành.

**(12) Thực thi suy đoán trong Hadoop là gì?**

Thực thi suy đoán là một kỹ thuật tối ưu hóa mà Hadoop sử dụng khi phát hiện DataNode đang thực hiện một tác vụ quá chậm. Có thể có nhiều lý do khiến tốc độ chậm lại và có thể khó xác định nguyên nhân thực sự của nó. Thay vì cố gắng chẩn đoán và khắc phục sự cố, Hadoop xác định tác vụ đang được đề cập và khởi chạy một tác vụ tương đương -- tác vụ suy đoán -- làm bản sao lưu. Nếu tác vụ gốc hoàn thành trước tác vụ suy đoán, Hadoop sẽ hủy tác vụ suy đoán đó.

**(13) Mục đích của lệnh JPS trong Hadoop là gì?**

JPS, viết tắt của Java Virtual Machine Process Status, là lệnh được sử dụng để kiểm tra trạng thái của các daemon Hadoop, cụ thể là NameNode, DataNode, ResourceManager và NodeManager. Người quản trị có thể sử dụng lệnh để xác minh xem các daemon có đang hoạt động hay không. Công cụ trả về ID quy trình và tên quy trình của mỗi Java Virtual Machine (JVM) đang chạy trên hệ thống đích.

**(14) Bạn có thể sử dụng lệnh nào để khởi động và dừng tất cả các daemon Hadoop cùng một lúc?**

Bạn có thể sử dụng lệnh sau để khởi động tất cả các daemon Hadoop:

./sbin/bắt đầu-tất-cả.sh

Bạn có thể sử dụng lệnh sau để dừng tất cả các daemon Hadoop:

./sbin/dừng-tất-cả.sh

Hình: Kiến trúc YARN của Hadoop

Kiến trúc của YARN.

**(15) Nút cạnh trong Hadoop là gì?**

Một nút biên là một máy tính hoạt động như một cổng thông tin người dùng cuối để giao tiếp với các nút khác trong cụm Hadoop. Một nút biên cung cấp một giao diện giữa cụm Hadoop và mạng bên ngoài. Vì lý do này, nó cũng được gọi là nút cổng hoặc nút giao tiếp biên. Các nút biên thường được sử dụng để chạy các công cụ quản trị hoặc ứng dụng khách. Chúng thường không chạy bất kỳ dịch vụ Hadoop nào.

**(16) Sự khác biệt chính giữa NFS và HDFS là gì?**

NFS , viết tắt của Network File System, là một giao thức hệ thống tệp phân tán được triển khai rộng rãi được sử dụng rộng rãi trong các hệ thống lưu trữ gắn mạng ( NAS ). Đây là một trong những hệ thống lưu trữ tệp phân tán lâu đời nhất và phù hợp với các tập dữ liệu nhỏ hơn. NAS giúp dữ liệu có sẵn qua mạng nhưng có thể truy cập như các tệp trên máy cục bộ.

HDFS là công nghệ mới hơn. Nó được thiết kế để xử lý khối lượng công việc dữ liệu lớn, cung cấp thông lượng cao và dung lượng lớn, vượt xa khả năng của hệ thống dựa trên NFS. HDFS cũng cung cấp các biện pháp bảo vệ dữ liệu tích hợp để bảo vệ chống lại lỗi nút. NFS thường được triển khai trên các hệ thống đơn lẻ không bao gồm khả năng chịu lỗi vốn có đi kèm với HDFS. Tuy nhiên, các hệ thống dựa trên NFS thường ít phức tạp hơn nhiều để triển khai và bảo trì so với các hệ thống dựa trên HDFS.

**(17) Phần cứng thương mại là gì?**

Phần cứng hàng hóa là thiết bị hoặc thành phần có sẵn rộng rãi, tương đối rẻ và thường có thể được sử dụng thay thế cho các thành phần khác . Phần cứng hàng hóa đôi khi được gọi là phần cứng có sẵn vì tính sẵn có và dễ mua. Các tổ chức thường chọn phần cứng hàng hóa thay vì phần cứng độc quyền vì nó rẻ hơn, đơn giản hơn và nhanh hơn để mua, và dễ thay thế tất cả hoặc một số thành phần trong trường hợp phần cứng bị hỏng. Phần cứng hàng hóa có thể bao gồm máy chủ, hệ thống lưu trữ, thiết bị mạng hoặc các thành phần khác.

**(18) MapReduce là gì?**

MapReduce là một khuôn khổ phần mềm trong Hadoop được sử dụng để xử lý các tập dữ liệu lớn trên một cụm máy tính trong đó mỗi nút bao gồm bộ lưu trữ riêng của nó. MapReduce có thể xử lý dữ liệu song song trên các nút này, giúp phân phối dữ liệu đầu vào và đối chiếu kết quả. Theo cách này, Hadoop có thể chạy các tác vụ được chia trên một số lượng lớn máy chủ. MapReduce cũng cung cấp mức độ chịu lỗi riêng, với mỗi nút báo cáo định kỳ trạng thái của nó cho một nút chính. Ngoài ra, MapReduce cung cấp hỗ trợ gốc để viết các ứng dụng Java, mặc dù bạn cũng có thể viết các ứng dụng MapReduce bằng các ngôn ngữ lập trình khác.

**(19) Hai giai đoạn chính của hoạt động MapReduce là gì?**

Hoạt động MapReduce có thể được chia thành hai giai đoạn chính sau :

Giai đoạn bản đồ. MapReduce xử lý dữ liệu đầu vào, chia thành các khối và ánh xạ các khối đó để chuẩn bị cho phân tích. MapReduce chạy các quy trình này song song.

Giảm pha. MapReduce xử lý các khối được ánh xạ, tổng hợp dữ liệu dựa trên logic đã xác định. Đầu ra của các pha này sau đó được ghi vào HDFS.

Các hoạt động MapReduce đôi khi được chia thành các giai đoạn khác ngoài hai giai đoạn này. Ví dụ, giai đoạn Reduce có thể được chia thành giai đoạn Shuffle và giai đoạn Reduce. Trong một số trường hợp, bạn cũng có thể thấy giai đoạn Combiner, đây là giai đoạn tùy chọn được sử dụng để tối ưu hóa các hoạt động MapReduce.

**(20) Lựa chọn tính năng trong dữ liệu lớn là gì?**

Lựa chọn tính năng đề cập đến quá trình trích xuất chỉ thông tin cụ thể từ một tập dữ liệu. Điều này có thể làm giảm lượng dữ liệu cần phân tích, đồng thời cải thiện chất lượng dữ liệu được sử dụng để phân tích. Lựa chọn tính năng giúp các nhà khoa học dữ liệu tinh chỉnh các biến đầu vào mà họ sử dụng để lập mô hình và phân tích dữ liệu, dẫn đến kết quả chính xác hơn, đồng thời giảm chi phí tính toán.

Các nhà khoa học dữ liệu sử dụng các thuật toán phức tạp để lựa chọn tính năng, thường thuộc một trong ba loại sau:

Phương pháp lọc. Một tập hợp con các biến đầu vào được chọn trong giai đoạn tiền xử lý bằng cách xếp hạng dữ liệu dựa trên các yếu tố như tầm quan trọng và mức độ liên quan.

Phương pháp bao bọc. Phương pháp này là hoạt động tốn nhiều tài nguyên, sử dụng máy học và phân tích dự đoán để cố gắng xác định biến đầu vào nào cần giữ lại, thường cung cấp kết quả tốt hơn so với phương pháp lọc.

Phương pháp nhúng. Phương pháp nhúng kết hợp các thuộc tính của cả phương pháp tệp và phương pháp wrapper, sử dụng ít tài nguyên tính toán hơn phương pháp wrapper, trong khi cung cấp kết quả tốt hơn phương pháp lọc. Tuy nhiên, phương pháp nhúng không phải lúc nào cũng hiệu quả bằng phương pháp wrapper.

**(21) "Giá trị ngoại lệ" trong bối cảnh dữ liệu lớn là gì?**

Một giá trị ngoại lệ là một điểm dữ liệu có khoảng cách bất thường so với các điểm khác trong một nhóm các mẫu ngẫu nhiên. Sự hiện diện của các giá trị ngoại lệ có khả năng gây hiểu lầm cho quá trình học máy và dẫn đến các mô hình không chính xác hoặc kết quả không đạt tiêu chuẩn. Trên thực tế, một giá trị ngoại lệ có khả năng làm sai lệch toàn bộ tập kết quả. Tuy nhiên, đôi khi các giá trị ngoại lệ có thể chứa đựng những thông tin giá trị.

**(22) Hai kỹ thuật phổ biến để phát hiện giá trị ngoại lai là gì?**

Các nhà phân tích thường sử dụng hai kỹ thuật sau để phát hiện các giá trị ngoại lệ:

Phân tích giá trị cực đại. Đây là dạng cơ bản nhất của phát hiện giá trị ngoại lai và chỉ giới hạn ở dữ liệu một chiều. Phân tích giá trị cực đại xác định chi tiết thống kê của phân phối dữ liệu. Điểm Z của Altman là một ví dụ điển hình về phân tích giá trị cực đại.

Các mô hình xác suất và thống kê. Các mô hình xác định các trường hợp không có khả năng xảy ra từ một mô hình dữ liệu xác suất. Các điểm dữ liệu có xác suất thành viên thấp được đánh dấu là các giá trị ngoại lệ. Tuy nhiên, các mô hình này giả định rằng dữ liệu tuân theo các phân phối cụ thể. Một ví dụ phổ biến về loại phát hiện giá trị ngoại lệ này là mô hình xác suất Bayesian .

Đây chỉ là hai trong số các phương pháp cốt lõi được sử dụng để phát hiện các giá trị ngoại lai. Các phương pháp tiếp cận khác bao gồm các mô hình hồi quy tuyến tính, các mô hình lý thuyết thông tin, các phương pháp phát hiện giá trị ngoại lai có chiều cao và các phương pháp tiếp cận khác.

**(23) Lệnh FSCK được sử dụng để làm gì?**

FSCK, viết tắt của kiểm tra tính nhất quán của hệ thống tệp, là tiện ích kiểm tra hệ thống tệp HDFS tạo báo cáo tóm tắt về trạng thái của hệ thống tệp. Tuy nhiên, báo cáo chỉ xác định sự hiện diện của lỗi; không sửa lỗi. Lệnh FSCK có thể được thực thi trên toàn bộ hệ thống hoặc một tập hợp con tệp được chọn.

Hình: YARN so với MapReduce

YARN và MapReduce hỗ trợ quản lý cụm Hadoop nhưng có cách tiếp cận khác nhau.

**(24) Bạn có sẵn sàng học thêm kiến ​​thức và nâng cao trình độ để giúp bạn thăng tiến sự nghiệp cùng chúng tôi không?**

Đây là cơ hội để bạn thể hiện sự nhiệt tình và tham vọng nghề nghiệp của mình. Tất nhiên, câu trả lời của bạn sẽ phụ thuộc vào trình độ học vấn và chứng chỉ hiện tại của bạn, cũng như hoàn cảnh cá nhân của bạn, có thể bao gồm trách nhiệm gia đình và cân nhắc về tài chính. Do đó, hãy trả lời một cách thẳng thắn và trung thực. Hãy nhớ rằng nhiều khóa học và mô-đun học tập có sẵn trực tuyến. Hơn nữa, các nhà cung cấp phân tích đã thiết lập các khóa đào tạo nhằm vào những người muốn nâng cao kỹ năng của mình trong lĩnh vực này. Bạn cũng có thể hỏi về chính sách cố vấn và huấn luyện của công ty.

**(25) Bạn có câu hỏi nào cho chúng tôi không?**

Như đã đề cập trước đó, một nguyên tắc chung là đi phỏng vấn với một vài câu hỏi đã chuẩn bị. Nhưng tùy thuộc vào cách cuộc trò chuyện diễn ra trong buổi phỏng vấn, bạn có thể chọn không hỏi chúng. Ví dụ, nếu chúng đã được trả lời hoặc cuộc thảo luận đã khơi dậy những câu hỏi khác có liên quan hơn trong đầu bạn, bạn có thể gác lại những câu hỏi ban đầu của mình. Ngoài ra, hãy lưu ý đến cách bạn đặt câu hỏi, hãy theo gợi ý của người phỏng vấn. Tùy thuộc vào hoàn cảnh, việc đặt câu hỏi trong buổi phỏng vấn là chấp nhận được, mặc dù nói chung, việc trì hoãn câu hỏi của bạn cho đến khi kết thúc buổi phỏng vấn là phổ biến hơn. Tuy nhiên, bạn không bao giờ nên ngần ngại yêu cầu làm rõ câu hỏi mà người phỏng vấn hỏi.