캘리포니아 대학교 버클리의 연구진은 정확하고 정밀한 의학 진단을 위하여 Apache Spark를 기반으로 구축된 최신 툴을 사용하여 DNA 염기서열 분석을 진행하고 있습니다. 퓨어스토리지 플래시블레이드(FlashBlade™)의 도입으로 대량의 DNA 샘플 데이터에 대한 염기서열 분석 및 그 결과를 바탕으로 한 의료정보 제공에 대한 시간이 대폭 감소되었습니다.

Berkeley
UNIVERSITY OF CALIFORNIA

비즈니스 혁신

연구진과 임상 의료진은 기존보다 적은 시간 동안 더 많은 DNA 샘플에 대한 염기서열 분석을 수행하여 통찰력을 확보하고, 생명이 위독한 환자에게 신속한 처방을 내릴 수 있게 되었습니다.

지역

북미

산업

헬스케어

"퓨어스토리지 플래시블레이드는 대역폭을 수평적으로 확장하여 높은 처리량을 유지할 수 있도록 해준다."

프랭크 어스틴 노세프트(Frank Austin Nothaft), 대학원생

캘리포니아 대학 버클리의 연구진은 APACHE SPARK과 퓨어스토리지의 플래시블레이드를 사용하여 유전체학의 경계를 확장했습니다.

인류가 인간 게놈을 배열할 수 있는 기술을 가짐으로써 역사상 드물게 인류에 대한 엄청난 잠재력을 지닌 과학적 돌파구가 마련되었습니다. 연구진과 임상 의료진은 수천여 질병의 예방과 치료에 유전체 염기서열 분석 결과를 활용합니다. 그 이유는, 유전자 정보를 활용하지 않는 기존 방식보다 더 빠르고 더 적은 비용으로 수천여 질병의 예방과 치료를 할 수 있게 되기 때문입니다.

고급 분석 툴이 없다면 DNA 염기서열 분석은 불가능할 것입니다. 그러나 이러한 툴의 기능이 지속적으로 향상되어도, 여전히 커다란 도전과제가 남아있습니다. 염기서열 분석은 방대한 데이터 처리가 필요한 작업이기 때문입니다. 유전체 염기서열 분석을 설명하기에 단지 '빅데이터'라는 용어는 턱없이 부족합니다. 한 사람의 DNA 샘플에는 약 300GB의 원시 염기서열 데이터가 존재하며 최신 DNA 염기서열 분석 연구 프로젝트에는 50,000~100,000명이 참여합니다. 또한, 연구조사 기간 중 각 참여자들로부터 수 차례 DNA 샘플을 채취하게 됩니다. 단일한 프로젝트의 데이터베이스만 해도 페타바이트 규모가 될 수 있습니다. 전세계적으로 이러한 거대 규모의 프로젝트가 다수 진행 중이거나 혹은 계획 중에 있습니다.

이렇게 많은 유전체 염기서열 분석이 진행되고 있는 큰 이유 중 하나는 그 비용이 획기적으로 감소되었기 때문입니다. 2003년 완료된 인간 게놈 프로젝트(Human Genome Project) 이후부터 현재까지, 기술의 발달로 사람 한 명의 유전체 염기서열 분석에 필요한 비용은 1억 달러에서 1천 달러로 급감했습니다.

그러나 유전체 염기서열 분석의 폭발적인 증가에는 또 다른 중요한 이유가 있습니다. 앞으로 인간의 질병에 대한 예방과 생명연장이라는 긍정적인 기대 때문입니다.

캘리포니아 대학교 버클리의 전기 엔지니어링 및 컴퓨터 과학부의 교수이자 버클리 컴퓨터 생명공학 센터(Center for Computational Biology at Berkeley)의 교수인 앤서니 조셉(Anthony Joseph)은 "오늘날의 의학은 의료 서비스를 제공하는데 있어 아직도 부족한 정보를 가지고 있다. 그러나, 만약 의학이 모든 사람에 대한 유전체 정보를 분석할 수 있게 되면, 개인에 맞춤화된 보다 정밀한 의학을 제공할 수 있게 된다."고 말합니다.

스케일-아웃 스토리지는 신속한 DNA 염기서열 분석에서 핵심적인 역할을 합니다.

프랭크 어스틴 노세프트(Frank Austin Nothaft)는 버클리 캠퍼스에 위치한 RISELab에서 찰스 치우 박사의 DNA 염기서열 분석 작업에 기반이 될 유전체 분석 툴인 고성능 오픈소스 분산형 라이브러리 ADAM을 개발하고 있습니다.

"2013년 시작된 ADAM의 개발 작업은 많은 염기서열 분석 툴이 계산적으로 효율적이지 않다는 중요한 단점을 해결하기 위한 것이다."고 노세프트는 설명합니다. 목표는 수천 줄의 코드를 사용해 실행하는데 한 달이 소요되는 프로세스를 단 100줄만을 사용해 하루에 실행하는 것입니다.

고객명:

캘리포니아 대학교 버클리
(University of California-Berkeley)
www.berkeley.edu

활용 사례:

- 데이터 분석 – Apache Spark®

도전 과제:

- 유전체 데이터의 폭발적 증가로 인해 고성능 스토리지에 대한 새로운 접근방식 필요
- 유전체 염기서열 분석 프로세스에 대한 재고 필요
- 스토리지 용량 추가시 비용 지출 과다 및 다운타임 발생

IT 혁신:

- 핵심적인 염기서열 분석 인덱스를 위한 로드 시간 3배 가속화
- 컴퓨팅 성능과 스토리지 성능의 분리를 통해 스토리지 용량 추가 간소화, 분석작업에 필요한 민첩성 확보 및 비용 절감
- 성능에 지장을 주지 않고 다양한 유형의 분석 작업들에 대한 동시적 지원이 가능해져 연구진에게 보다 폭넓은 과제를 탐구하는데 핵심적인 유연성 제공

"비트 매칭 분석은 유전체 분석 프로세싱에서 매우 중요하다. 플래시블레이드 외에 이를 수행할 수 있는 다른 방법은 존재하지 않는다."

프랭크 어스틴 노세프트(Frank Austin Nothaft), 대학원생

"RISELab팀은 유전체 분석 툴 개발에 기존 방식과는 다른 새로운 접근을 시도하였고, RISELab팀은 고성능 분산형 데이터베이스의 형태를 한 클러스터 구조를 선택했다." ADAM은 Apache Spark 오픈소스 프레임워크에 기반한 유전체 분석 툴 입니다.

ADAM을 위한 IT 인프라를 설계하는 과정에서, 데이터 저장방식을 스케일-아웃 방식으로 결정하였으나 실제로 이것을 구현하는 데는 여러가지 어려움이 있었습니다. 노세프트는 "유전체 분석 과정에서 원시 유전체 정보를 읽고 조합하는 과정이 대량으로 발생하므로 스토리지의 중요성은 매우 크다"고 말합니다.

ADAM의 초기 버전 스토리지 인프라는 HDD 디스크와 HDFS(하둡 파일 시스템)을 사용하였습니다. 기존 스토리지 아키텍처는 유전체 분석 툴 개선을 위한 고성능 스토리지 아키텍처에 대한 요구조건을 충족시키지 못했습니다.

"스토리지 성능에 대한 요구가 컴퓨팅 성능에 대한 요구보다 더 컸다. ADAM의 HDFS 성능 분석 결과, 평균 10~30%의 컴퓨팅 자원을 사용하는 반면, 스토리지 자원(용량, I/O)은 80~85%를 사용했다. 따라서 용량 확장과 성능 확장을 동시에 갖춘 스토리지 인프라가 필요했다"고 노세프트는 설명했습니다.

이제 유전체 분석 기술의 도움으로 위스콘신 주 코티지 그로브에 사는 15세 소년 조슈아 오스본(Joshua Osborn)이 어떻게 위중한 상태에서 건강한 모습으로 가족의 품으로 돌아가게 되었는지 설명 드리겠습니다. 조슈아는 뇌염(뇌 부종) 진단을 받았습니다. 증세가 심각하여 약물을 사용해 인위적인 혼수상태에 들어갔습니다. 다양한 테스트를 시도했지만 조슈아의 상태에 대해 정확한 원인을 파악할 수 없었습니다.

최종적으로, 샌프란시스코의 캘리포니아 대학교에 있는 찰스 치우(Charles Chiu) 박사에게 조슈아의 뇌염의 원인에 대한 분석 의뢰가 들어왔습니다. 치우 박사는 DNA 염기서열 분석 결과를 신속하게 판독 및 분석하여, 이를 질병을 야기하는 병원체와 일치시키는 방법을 개발해왔습니다. 뇌염의 원인을 빨리 찾기 위하여, 치우 박사는 한번에 수백만 건의 시퀀스를 생성하는 최첨단 DNA 염기서열 분석 툴을 사용했습니다. [1]

조슈아의 혈액 및 척수액을 전달 받은 지 48시간 내에, 연구소의 과학자들은 3백만 개의 DNA 염기서열에 대한 분석을 완료하고 원인을 밝혀냈습니다. 뇌염의 원인은 조슈아와 가족이 방문한 적이 있는 푸에르토리코에서 서식하는 박테리아균이었습니다. 이러한 진단을 바탕으로, 조슈아의 의료진은 조슈아에게 페니실린을 처방했습니다. 조슈아는 혼수 상태에서 깨어났으며 병세가 빠르게 호전되어 입원 76일만에 퇴원할 수 있었습니다.

치우 박사와 팀원들이 신속하게 염기서열 분석 결과를 확보할 수 있도록 해준 유전체 분석 툴은 조셉 박사의 지휘 하에 UC 버클리 팀이 개발했습니다. 조셉 박사는 "사람들이 유전체 데이터를 처리하고 분석할 수 있는 시스템을 구축하는데 중점을 둔다"며 "수집되는 데이터의 양이 폭발적으로 늘어나고 있으며, 해마다 더 많은 사람들의 유전체 염기서열을 분석하고 있다. 분석 비용은 획기적으로 감소하고 있으며, 유전체 분석의 영역 역시 향상되어 더 많은 분석자료를 수집할 수 있게 되었다"고 설명했습니다.

모든 핵심 요구사항을 충족하는 플래시블레이드

노세프트는 "현재의 클러스터 구성으로 HDFS를 중심으로 구축된 스토리지 용량을 늘리려는 시도는 무리가 있었다"며 "HDFS노드에 이미 가장 큰 용량의 디스크를 장착하고 있었고, 많은 비용을 들여 노드를 늘리는 것만이 유일한 옵션이었다"고 말했습니다.

RISELab은 퓨어스토리지 플래시블레이드 베타 사이트로서 플래시블레이드의 가능성을 확인 할 수 있었습니다. 퓨어스토리지의 플래시블레이드는 제품 설계 단계부터 대용량 병렬처리를 염두한 아키텍처로서 데이터 분석을 효과적으로 지원하는 진정한 올플래시 스토리지입니다.

1. 조슈아 사례 상세 정보 https://www.ucsf.edu/news/2014/06/115116/teen's-recovery-shows-value

"전에는 실행할 수 없었던 워크로드도 플래시블레이드로 문제 없이 실행하고 있다. 우리 팀에게는 엄청난 일이다."

프랭크 어스틴 노세프트(Frank Austin Nothaft), 대학원생

플래시블레이드는 대용량 데이터 스토어를 빠르게 처리할 수 있는 병렬 아키텍처, IOPS, 무한한 용량 확장성, 획기적으로 간소화 된 관리 등 스토리지 솔루션에 대한 UC 버클리의 모든 요구 조건을 충족했습니다.

노세프트는 "플래시블레이드는 간단하게 블레이드만 추가하면 수평적으로 대역폭을 확장할 수가 있으며 높은 처리량을 유지할 수 있도록 해준다"고 설명했습니다. "또한 클러스터에서 컴퓨팅 자원과 스토리지 자원을 분리하여 운영할 수 있게 해주었다. 기존의 HDFS는 성능확보를 위하여 항상 컴퓨팅 자원과 스토리지 자원을 동시에 구매해야 했다. 복잡한 계획 수립 단계도 필요하지 않고 운영에 지장을 주지도 않는다. 유전체학에서 이는 대단한 일이다. 유전체 분석에 높은 성능도 필요하며, 동시에 대용량의 저장소도 필요하다."고 말했습니다.

플래시블레이드 도입으로 성능 및 유연성 향상

ADAM에 플래시블레이드를 도입한 이후, 버클리 대학팀은 다양한 혜택을 얻었습니다. 노세프트는 " 플래시블레이드를 사용하면 유전체 염기서열 분석을 가속화할 수 있다. 성능, 관리편의성, 확장성을 향상시켜주기 때문이다"며 "플래시블레이드를 이용하여 유전체 분석을 하면서, 분석 성능이 3배나 향상되었으며, 처리량과 응답시간 또한 대폭 향상되었다"고 말했습니다.

유전체 염기서열 분석 단계 중 하나는 variant calling입니다. 플래시 블레이드가 구성된 ADAM의 variant calling 분석 성능은 17배 향상하였으며 비용은 절반으로 줄었습니다.

(variant calling은 각 개인의 염기서열차이를 분석하는 작업입니다. 인간의 염기서열은 30억개로 이루어져 있습니다. 여기서 99.9%는 모든 인간이 동일한 유전자 정보를 가지고 있고 0.1%, 300만개의 염기서열이 개인별로 차이를 가지고 있습니다. variant calling 분석을 통하여 개인별 의학 진단을 할 수 있는 개인화된 유전 정보를 제공합니다.)

유전체 정렬의 또 다른 핵심적인 단계는 유전체 염기서열을 설명하는 대용량(5~40GB) 인덱스를 로드하는 것입니다. HDFS로부터 마운트하는 대신 플래시블레이드로부터 인덱스를 로드함으로써 기존에 30분 이상 소요되던 정렬 작업을 11분만에 완료할 수 있게 되었습니다.

"유전자 분석 프로세스들 중에서 중요한 프로세스의 하나인 '비트 매칭(bit matching)'에서 우리는 스토리지에서 1억 개의 파일을 처리하려 하였다. 처음 이를 시도했을 때 HDFS에서는 오류가 발생하여 작업이 실패하였다. 그런데 플래시블레이드 도입 후에는 문제없이 작동됐다. 데이터 플랫폼으로서, 플래시블레이드는 엄청난 성능을 발휘한다. 굉장히 의미 있는 부분이다. 비트 매칭 분석은 유전체 프로세싱에 매우 중요하기 때문이다. 플래시블레이드 외에 이를 수행할 수 있는 다른 방법은 존재하지 않는다"고 노세프트는 덧붙였습니다.

또 다른 혜택은 유연성입니다. 그는 "대용량 스캔과 시퀀셜 액세스나 포인트 파일 검색처럼 랜덤 액세스가 동시에 발생하는 하이브리드 워크플로우를 다루는 경우가 종종 있다. 플래시블레이드는 성능 손실이나 재코딩 없이 이 두가지를 동시에 처리할 수 있도록 한다. 이를 통해 즉각적인 분석 역량을 향상시킬 수 있었다"고 말했습니다.

관리 편의성이 제공한 향상된 유전자 분석 연구 집중도

연구 자체와 ADAM 개발 지원에 집중하다 보면 시스템 관리 작업에 할애할 시간이 없습니다. " 플래시블레이드는 매우 간단하고 쉽게 사용할 수 있는 보고 인터페이스를 갖추고 있다. 원하는 통계 정보를 신속하게 얻을 수 있다"고 노세프트는 말했습니다. "플래시블레이드는 높은 안정성을 제공하도록 설계되었으며, 플레시 블레이드에 이슈가 발생하면, 퓨어스토리지의 퓨어1(Pure1)에서 이슈를 인지 및 분석하고,퓨어스토리지 지원팀이 해결방안에 대해 이메일로 알려준다. 관리의 측면에서 이는 큰 혜택이다."

ADAM 팀은 유전체 염기서열 분석을 통하여 새로운 발견을 가속화해 줄 개선점을 찾고 있습니다. 플래시블레이드를 통해 노세프트는 조슈아의 경우와 같은 성공 사례에 더 많이 기여할 수 있을 것입니다.

"연구자에게는 현재 실행할 수 없는 워크로드를 열어주는 모든 것이 소중하다. 전에는 실행할 수 없었던 일부 워크로드를 플래시블레이드 도입으로 문제 없이 실행할 수 있게 되었다. 이것은 우리 팀에게 대단한 성과이며, 나아가 우리가 미래에 긍정적 영향을 줄 수 있는 의학뿐만 아니라 다른 바이오 산업에게는 더 중대한 성과다"

PURESTORAGE®

korea@purestorage.com
www.purestorage.com/kr/customers

ps_cs_Univ-California-Berkeley_01