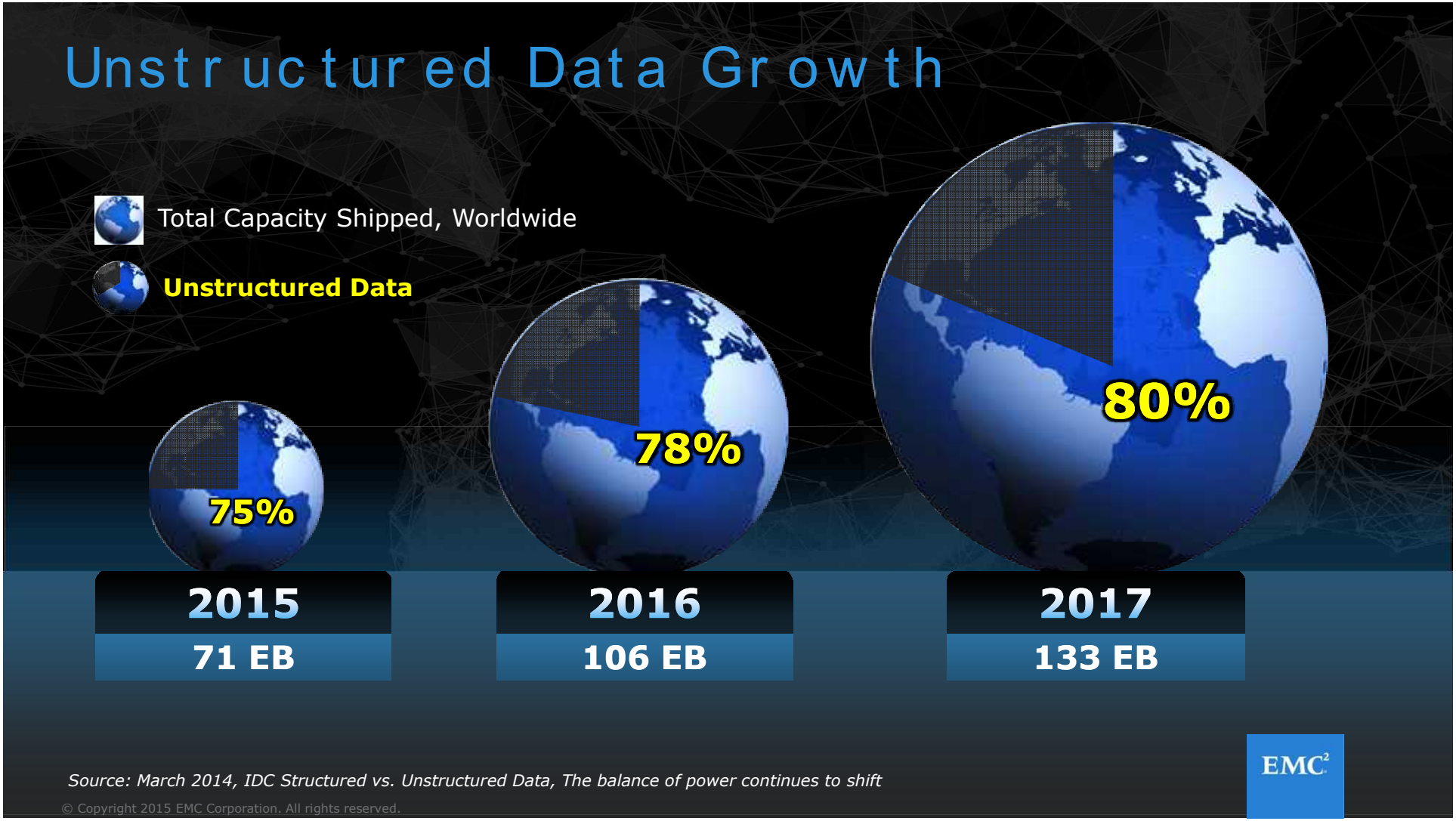
Scale-Out Data Lake Foundation

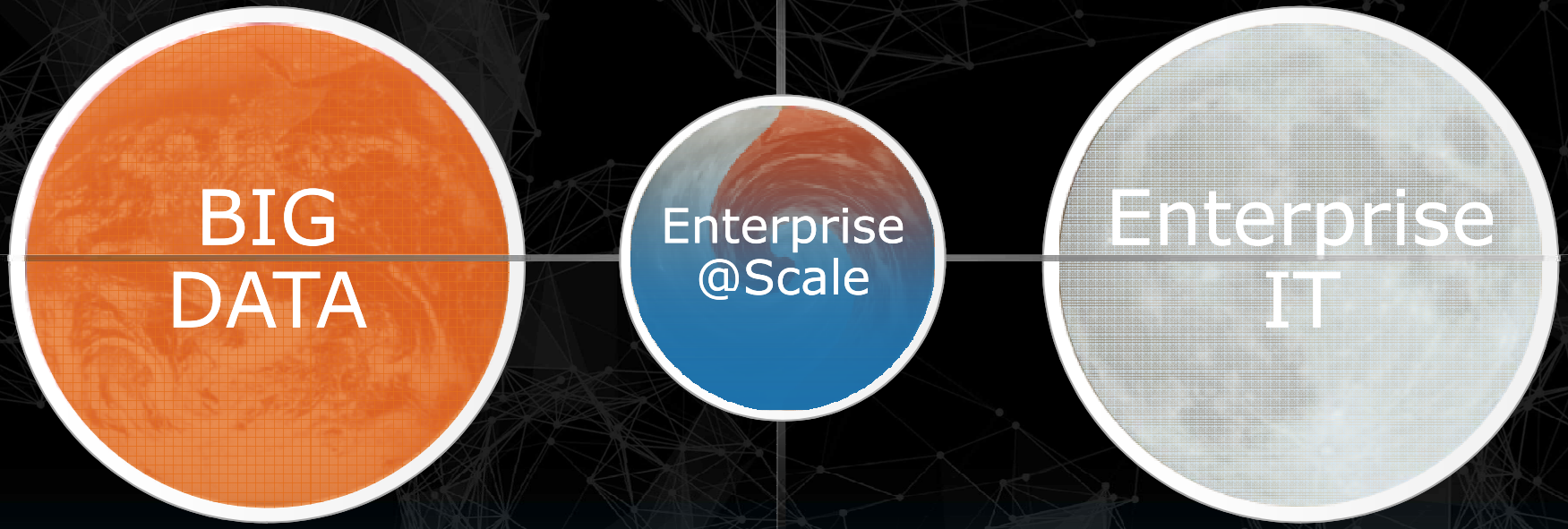
이 상우, Isilon 사업본부
한국 EMC

Unstructured Data Growth

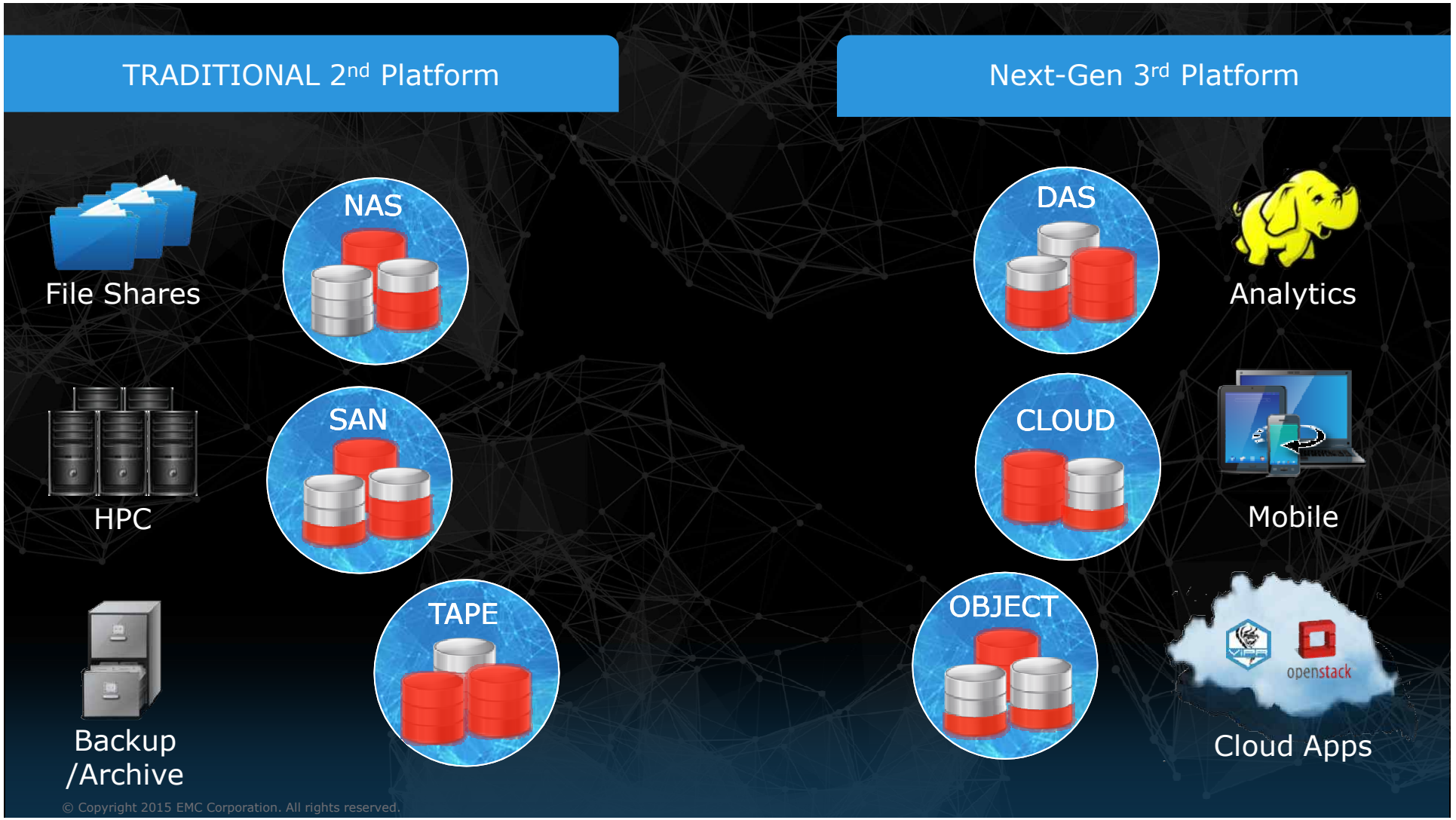


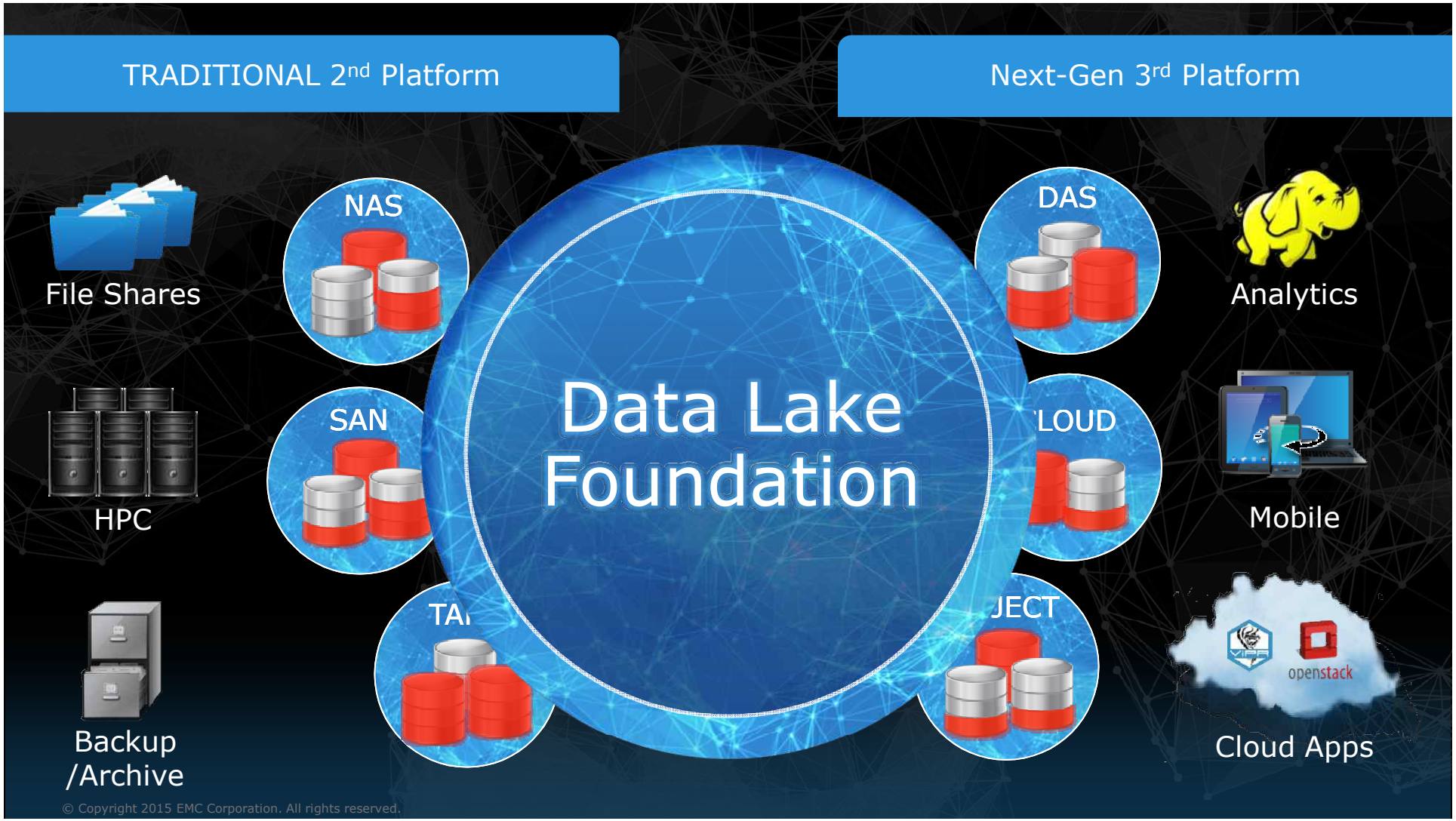
Source: March 2014, IDC Structured vs. Unstructured Data, The balance of power continues to shift

Two Storage Worlds Converge…

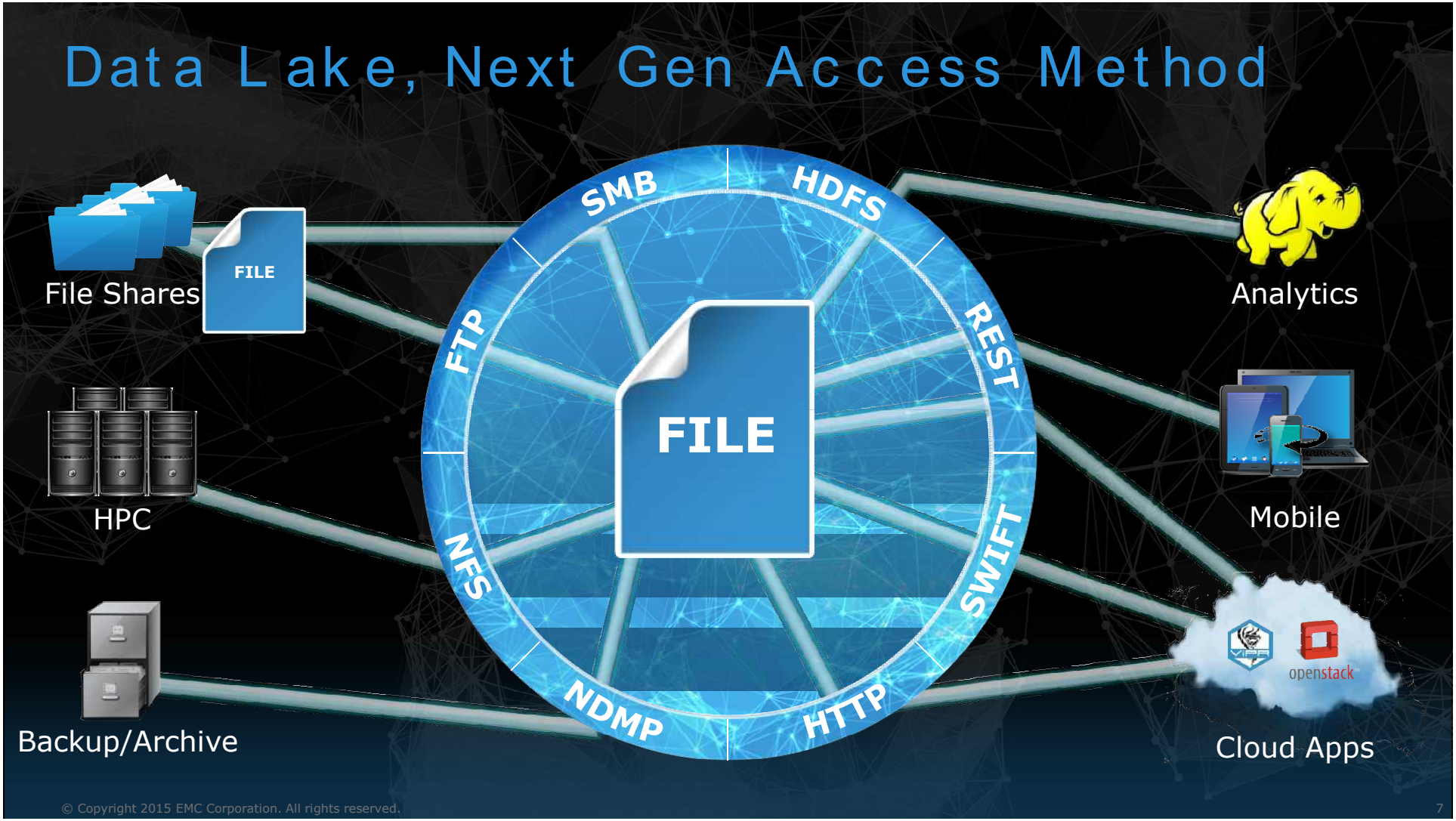


File Data Storage Needs Are Evolving





Data Lake, Next Gen Access Method



What is Data Lake?

- Single Unified Data Pool = "Single Copy of Truth"
- Unstructured & Semi-structured Data
- Multiple Access Points & Method
- Massive, Efficient Scalability
- Enterprise Data Governance & Protection
- Data Migration Immune

See IDC's Insight:
"Enterprise Data Lake Platforms: Deep Storage for Big Data & Analytics" July 2014

EMC²

Why need Data Lake?

- Too many application, data and infrastructure silos
  - 관리의 어려움, 확장에 대한 비용 부담
  - 만연해 있는 중복 데이터 (65% of capacity*IDC)
  - 통합된 데이터 공유와 Mobility 구성 환경이 필요
- Desire to leverage next generation analytics
  - 완벽한 데이터에 대한 수집 및 분석 능력
  - BI에서 "Data Science"(Predictive/Prescriptive)로의 이동
- Investment Protection During Rapid Change
  - 변화속에서도 Business를 위한 지속적인 운영 유지

EMC²

Data Lake Foundation

Data Lake 구현을 위한 근본적인 Data Storage Infrastructure

- Efficient Storage
  - Storage 고립을 제거, 편리한 Storage 관리, Storage 효율성
- Massive Scalability
  - Scale-Out 아키텍처, 쉽고 효과적인 관리와 동시에 엄청난 확장성 제공
- Increased Operational Flexibility
  - Multi Protocol, 광범위한 Traditional & next Gen 워크로드를 지원
- In-Place Big Data Analytics
  - Shared Storage Infrastructure, 빠른 결과 및 뛰어난 성능 제공
- Robust Data Protection and Security
  - 효과적이고 복원이 빠른 백업, 재해복구, 보안 옵션을 통한 데이터 자원 보호

Scale-Out Data Lake Foundation

Requirements

- ✓ 멀티 워크로드와 애플리케이션 지원 (Traditional and emerging)
- ✓ Multi-protocol 지원
- ✓ 성능을 동반한 용량 확장 제공 (Scale-Out)
- ✓ 비용 절감을 위한 효과적이고 쉬운 관리 (단일 파일시스템)
- ✓ 강력한 데이터 분석 능력 지원
- ✓ 엔터프라이즈급 데이터 보호
- ✓ 보안 및 컴플라이언스 요구 조건 만족

BENEFITS

- ✓ 비효율적인 스토리지 고립화 제거
- ✓ 정보 공유 극대화
- ✓ 데이터 분석 가속화
- ✓ 데이터를 통한 의사 결정 제공
- ✓ 관리 간소화 및 비용 절감

Isilon, Scale-Out Data Lake Foundation

50PB 단일 볼륨/ 파일시스템

Multi-Protocol 파일, 오브젝트, Hadoop

Gartner IDC Scale-Out NAS leader

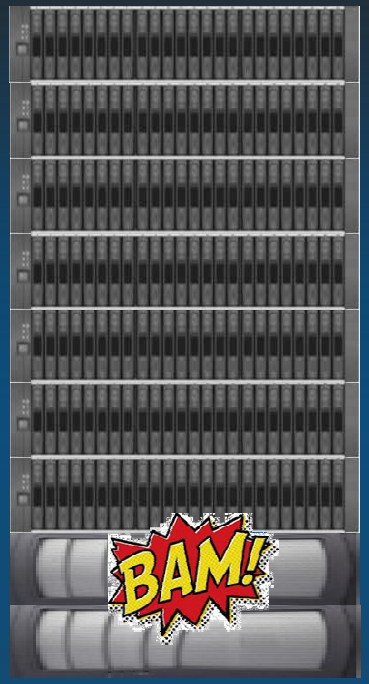
#1 Market Leader in Hadoop Shared Storage

600+ Analytics Customers

Pivotal cloudera Ask Bigger Questions Hortonworks

Scale-Out vs. Scale-Up

12.5%
성능감소
(8 노드 구성 시)



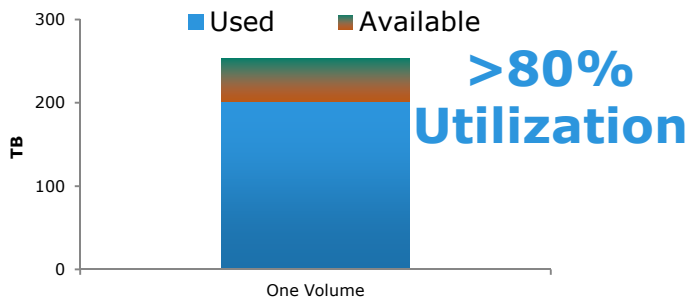
50%
성능감소

Scale-Out

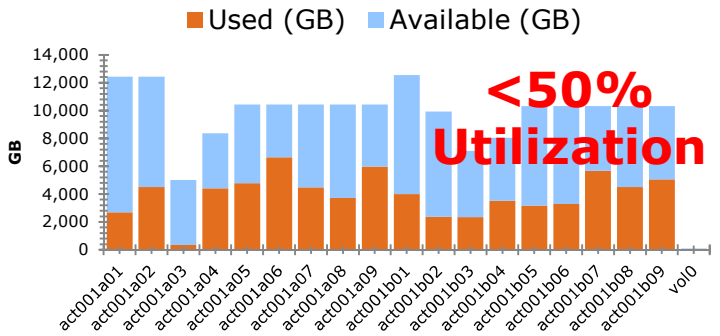
Scale-Up

단일 파일시스템 vs. 멀티 파일시스템

단일 파일시스템



멀티 파일시스템



EMC²

손쉬운 스토리지 확장

고성능 제공을 위한 자동 로드 밸런싱

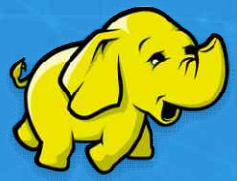
60초 이내

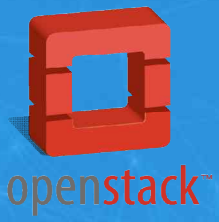
손쉬운 온라인 증설

최대

144 노드 / 50PB

CLOUD SCALE





NEXT GEN ACCESS METHODS

Scale-Out Data Lake Foundation

DATA MANAGEMENT

DATA PROTECTION

DATA SECURITY

Comprehensive Enterprise Software

DATA MANAGEMENT

가시성, 분석

InsightIQ

자동 계층화

SmartPools

저장 공간 효율화

SmartDedupe

Thin Provisioning

SmartQuotas

DATA PROTECTION

SyncIQ

D2D 백업, DR 복제

SnapshotIQ

논리적 백업

SmartLock

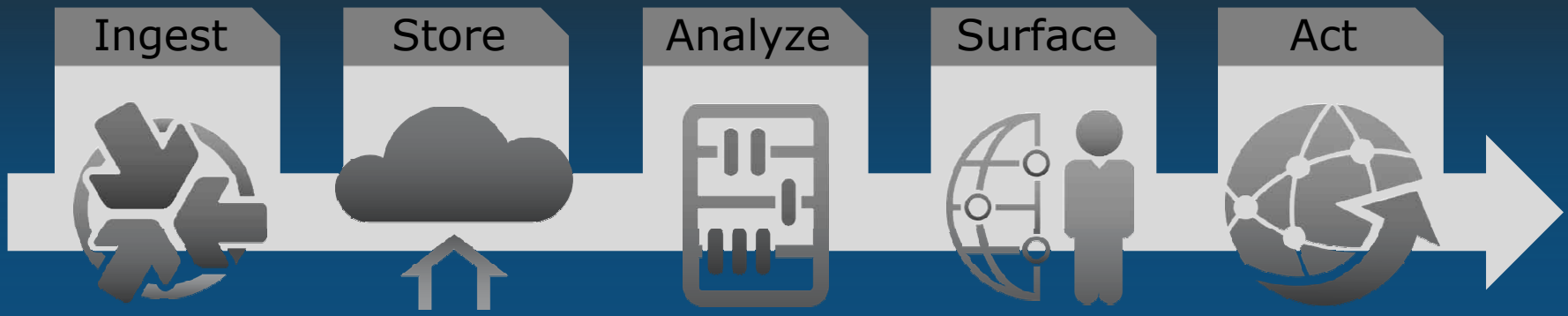
원본 위변조 방지

SmartConnect

커넥션 로드밸런싱

EMC²

Scale-Out Data Lake, 효과적인 분석 제공



광범위한 소스를 통한 자료 수집 (Traditional and new)

Data Set간 상호 분석을 위하여 단일 환경에 모두 저장

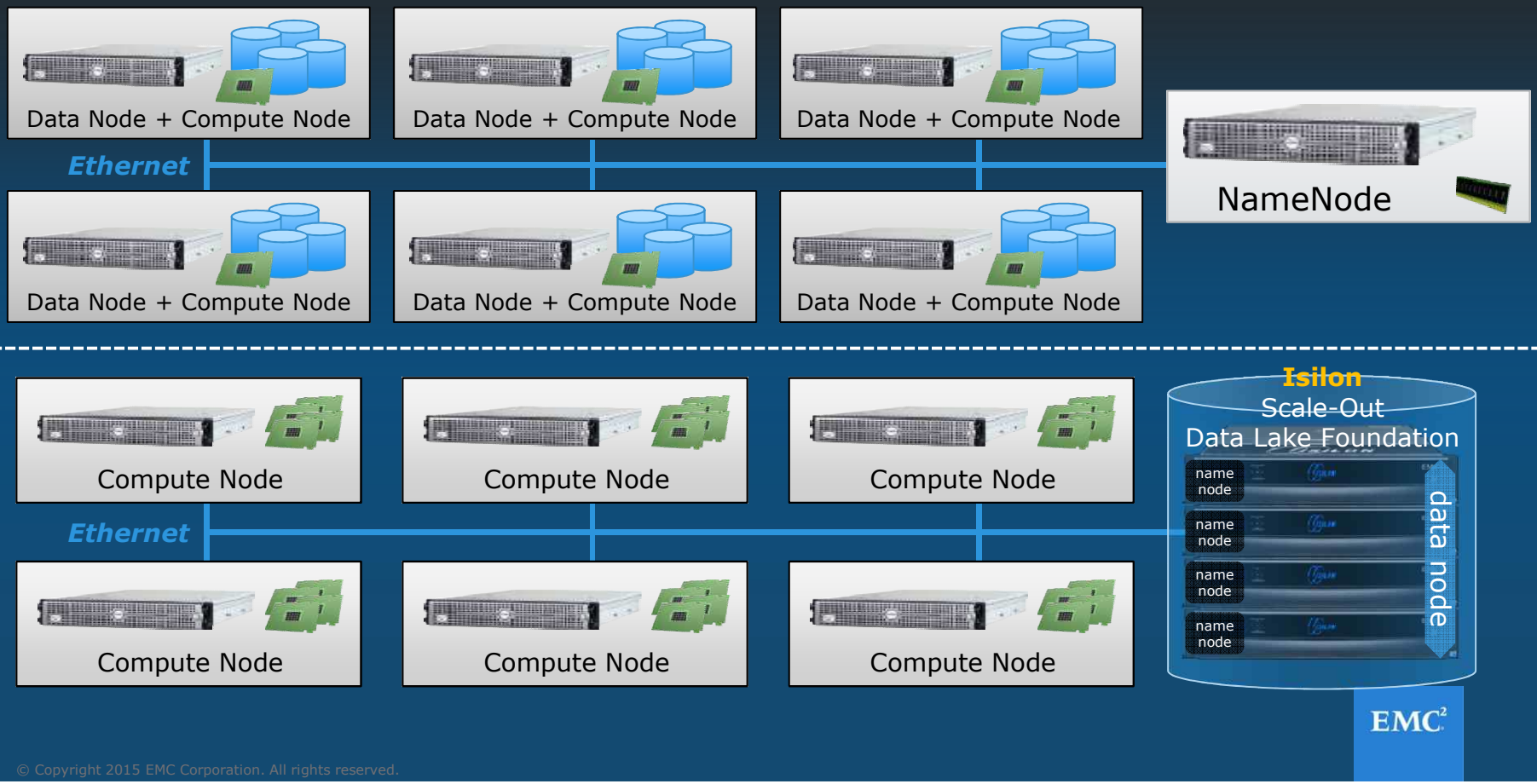
새로운, 미래 예측 패턴을 찾아내기 위하여 다양하고 진보된 알고리즘 사용

비즈니스 영역별 전문가들과 insight 를 공유

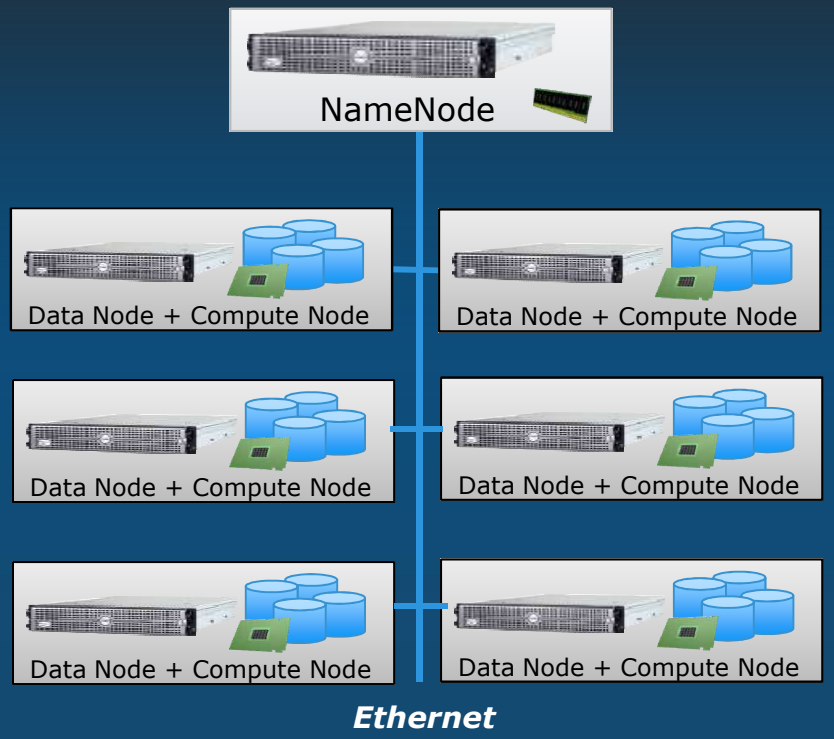
비즈니스 요구를 만족시키는 결과 지향적 애플리케이션 구축

EMC²

Hadoop 아키텍처 – DAS vs Isilon Scale-Out



Traditional Hadoop 아키텍처 - DAS

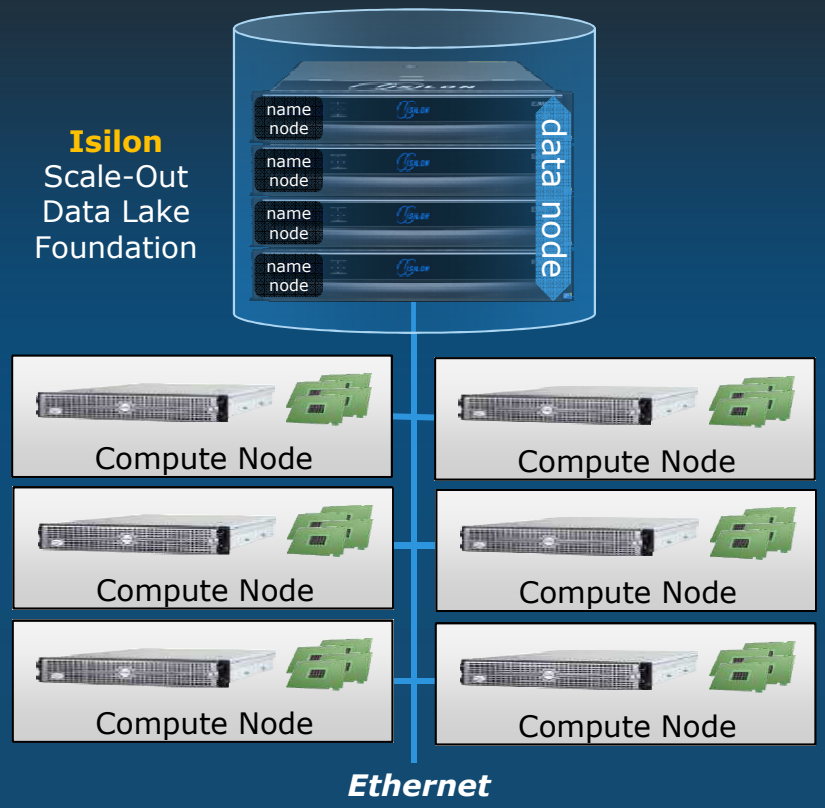


1	전용의 Storage Infrastructure – 오직 Hadoop만을 위해 사용
2	불완전한 HA 및 확장성 – Single Active NameNode
3	낮은 Enterprise Data 규정 준수 – No Snaps, DR, backup, WORM, DARE
4	낮은 Storage 효율성 – 3X 미러링
5	고정된 확장성 – Compute 노드와 Storage 노드를 함께 증설
6	Data Import/Export를 위한 매뉴얼 작업 – NAS protocol 지원 안됨

EMC²

Isilon Scale-Out Data Lake for Hadoop

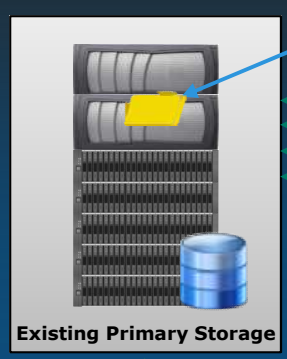
Isilon
Scale-Out
Data Lake
Foundation



1	멀티 프로토콜 Scale-Out Storage Platform – NFS, CIFS, FTP, HTTP, HDFS, RESTAPI
2	탄력적이고 예측 가능한 확장성 – 분산 NameNode & DataNode
3	엔터프라이즈 Data 보호 및 규정 준수 – SnapshotIQ, SyncIQ, SmartLock, ACLs..
4	업계 선도의 Storage 효율성 (단일 파일 시스템) – >80% Storage Utilization
5	독립적인 확장성으로 시스템 최적화 – 최적의 Storage 및 Compute 확장
6	고립된 데이터를 통합 – 업계 표준 프로토콜 – Shared Data를 통한 Application 실행

EMC²

전통적인 "Shar e-Not hing" Hadoop 구성



Existing Primary Storage

비정형 데이터

Primary Storage의 파일이 변경될 경우
데이터 정합성을 어떻게 유지하시겠습니까?

Existing Virtualized Data Center

EMC²

Isilon "Share-Everything" Hadoop

- 가상화 환경에서 유휴 자원으로 Hadoop 환경 구성
- 데이터 복제 불필요
  (기존 데이터 그대로 사용)
- 파일 프로토콜과 HDFS 프로토콜을 통해 동일 데이터를 동시에 액세스
- 별도의 데이터 복제본을 생성하지 않고 기존의 데이터를 바로 사용함으로써 비용절감 및 분석 기간을 획기적으로 당김

EMC²

Time-to-Results

DAS

Data Center Network

데이터 복제 | 분석

Primary Storage에서 Hadoop 시스템으로 100TB를 복사해 보신적인 있으신가요?

10GB 네트워크에서 100TB를 다른 곳으로 복사하는데 얼마나 많은 시간이 필요할까요?

>24 Hours

Isilon
Scale-Out
Data Lake Foundation

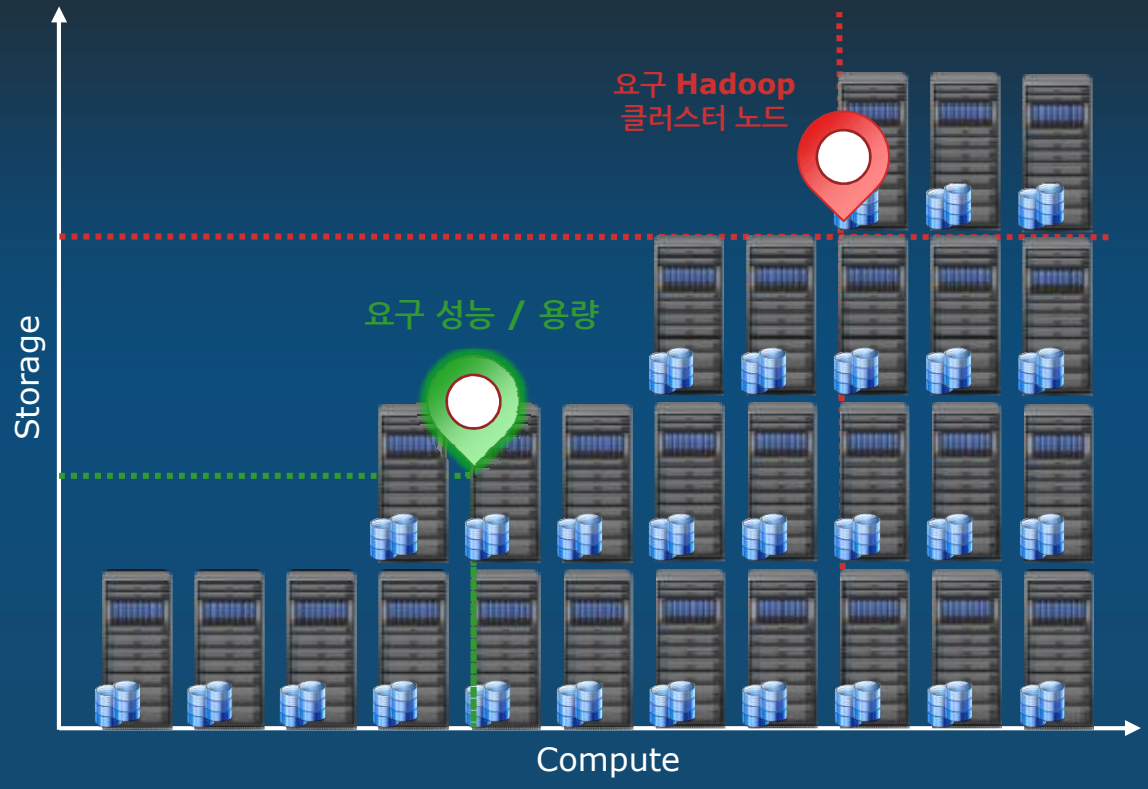
Data Center Network

즉시 분석

분석을 위해 관련 데이트를 직접 Read

EMC²

DAS
- 종속적 Scaling

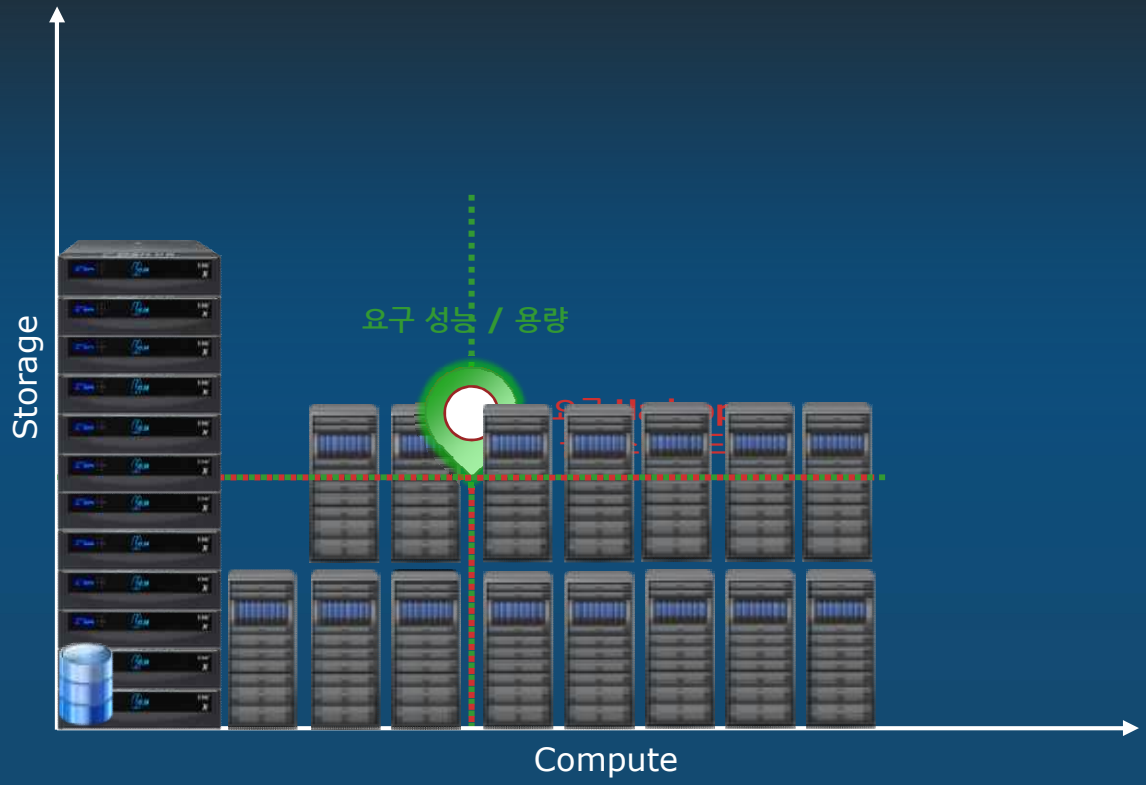


전통적 Hadoop HDFS

- Storage와 Compute의 정해진 비율
- Compute와 Storage가 함께 확장
- Compute 업그레이드는 불필요한 storage 부분까지 과도한 비용 발생

EMC²

Isilon, Scale-Out Data Lake
- 독립적 Scaling



전통적 Hadoop HDFS

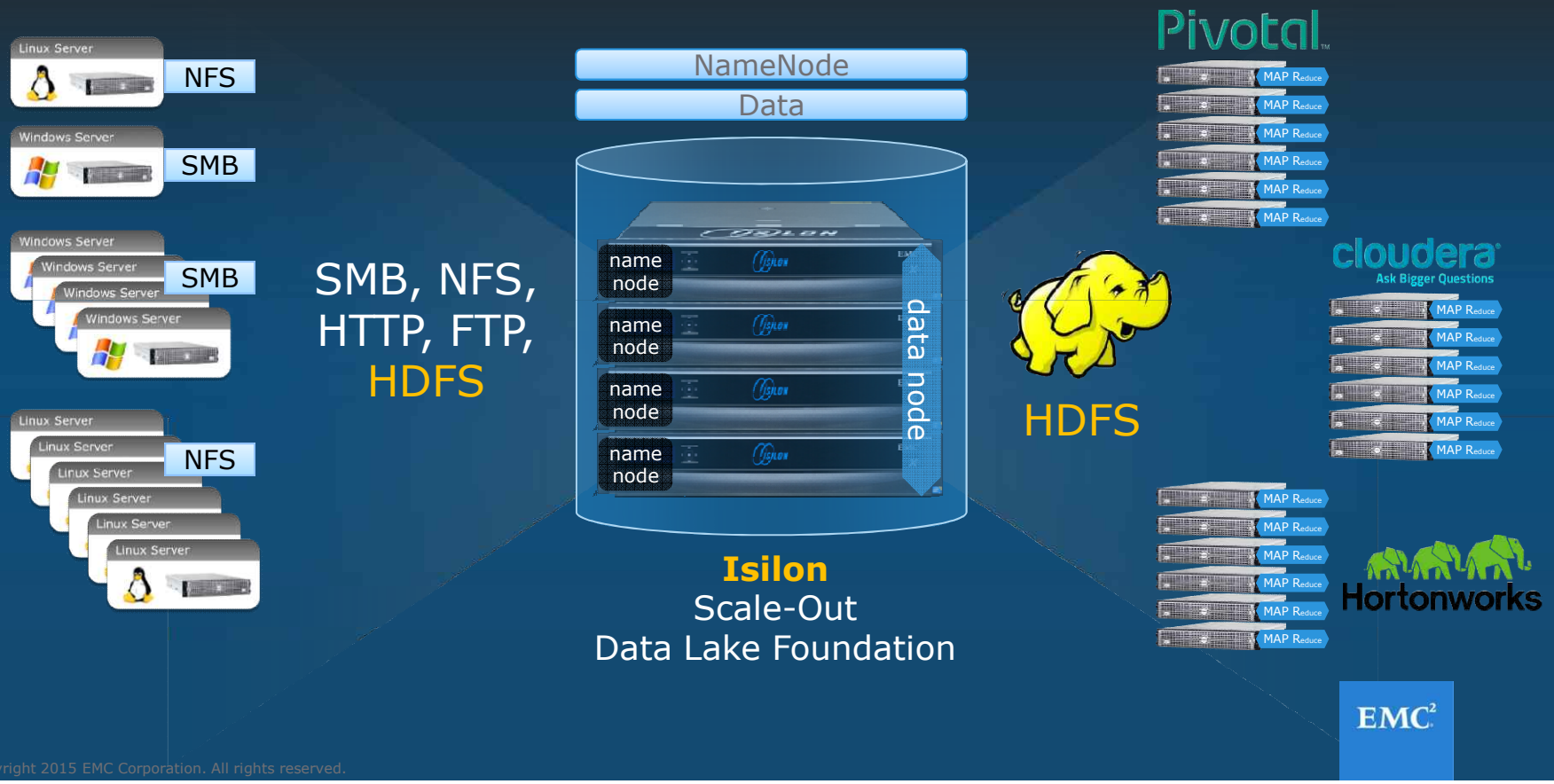
- Storage와 Compute의 정해진 비율
- Compute와 Storage가 함께 확장
- Compute 업그레이드는 불필요한 storage 부분까지 과도한 비용 발생

Isilon HDFS

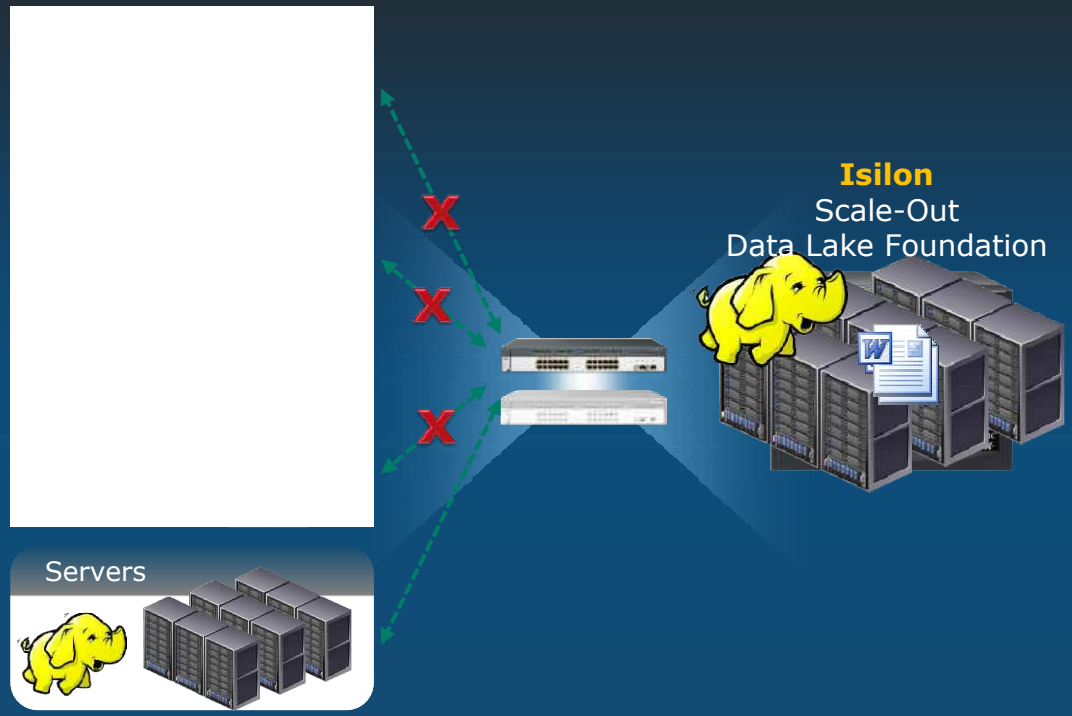
- Compute 부분만 독립적으로 확장하여 투자 비용에 대한 절감 효과
- 워크로드가 증가해도 이상적인 성능 밸런스를 이룸
- No data migrations, ever!
- 용량이 증가하면서 성능도 함께 증가

EMC²

Multiple Hadoop 분산 지원



Protocol 지원



Before with DAS

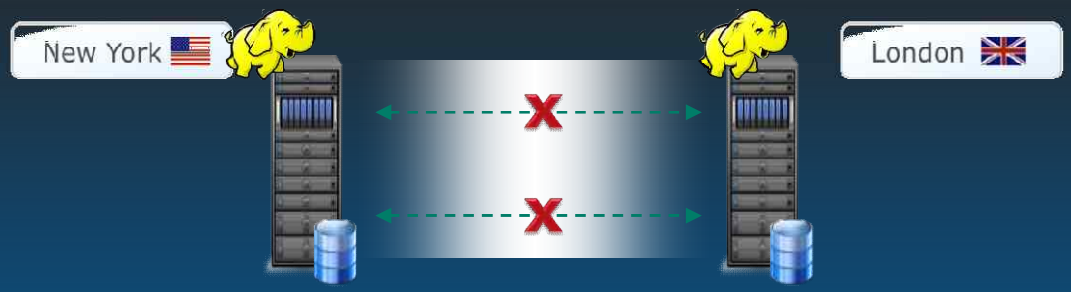
- HDFS는 Windows, Unix, Linux, Apple 혹은 다른 파일 시스템에서는 접근이 불가
- Big Data는 오직 Big Data용으로만…

After with Isilon

- Isilon 고유의 멀티 프로토콜 지원은 Hadoop을 포함 모든 파일시스템으로의 액세스를 가능하게 함.
- Big Data is actual data!

EMC²

Snapshot & Version Control



Before with DAS

- 전통적 HDFS은 replication 기능 부재
- No Snapshot for data
- 버전 관리 불가
- Mission Critical 상황에 대응 못함

Isilon
Scale-Out
Data Lake Foundation

Isilon
Scale-Out
Data Lake Foundation

After with Isilon

- SnapshotIQ를 통한 버전 관리
- Multi-threaded, Multi-Node Scale-Out replication 지원
- 업무 연속성을 위한 RPO/RTO 향상
- Geo-replicated Hadoop!

EMC²

Hadoop on Isilon,
Scale-Out Data Lake Foundation

- Hadoop 투자 비용을 줄일 수 있다. 
- Hadoop 구축이 쉽다 
- Hadoop 기존 데이터를 그대로 사용할 수 있다. 
- Hadoop NameNode 데이터도 protection이 가능하다. 
- Hadoop 데이터는 항상 최신 데이터로 유지할 수 있다. 
- Isilon은 Hadoop의 많은 문제점을 해결할 수 있다. 

EMC²

DAS 구성과 Isilon Hadoop 구성 비교

일반 Hadoop		Isilon Hadoop
원본 데이터 외에 3X 미러링의 분석용 데이터 필요	분석용 데이터	별도의 분석용 데이터 불필요, 원본 데이터를 그대로 사용
Hadoop 시스템으로 데이터 복사 후 분석 (많은 시간 필요)	**Time-To-Result**	원본 데이터를 이용, 즉시 분석 (빠른 분석 결과 도출)
분석을 위한 데이터를 다시 복사	원본 데이터 변경 시	원본 데이터를 사용하기 때문에 별도의 작업이 필요 없음
HDFS only	**Protocol**	**NFS, CIFS, HDFS, OBJECT** 동시 지원
Compute와 Storage의 종속적 확장 (불필요하고 과도한 비용 발생)	**Scalability**	Compute와 Storage의 독립적 확장 (투자 비용 절감)
별도의 Hadoop 시스템 구축	**Multi Hadoop**	**Multi Hadoop** 을 동시 지원
불가	**Replication & DR**	지원

Traditional Analytics Challenges

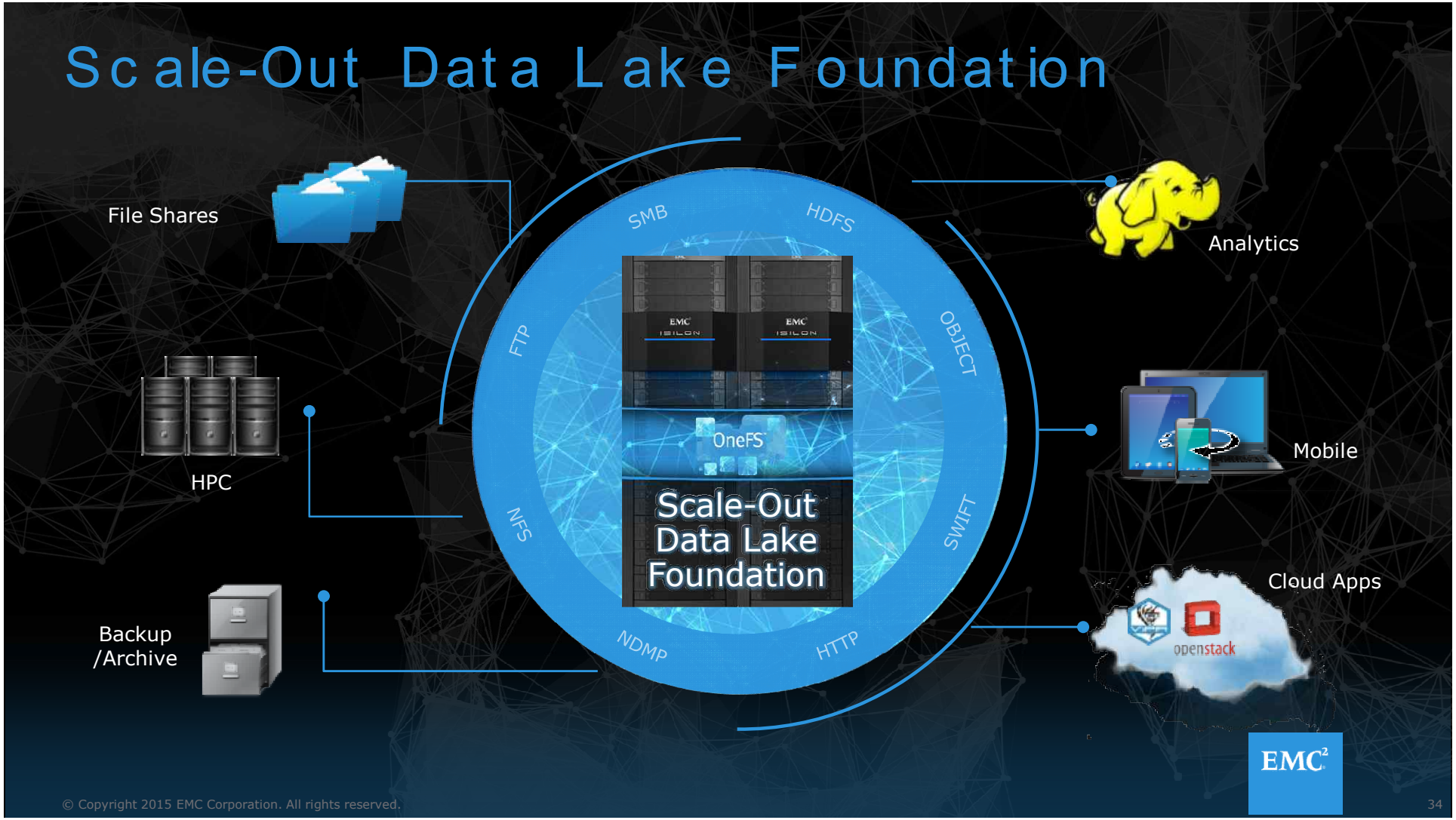
- 비정형 데이터 통합을 위한 실질적인 솔루션으로 보기 어려움
- 중복 데이터로 인한 고비용의 인프라스트럭처
- 고비용의 인프라스트럭처 및 소프트웨어
- 분석 시스템으로의 데이터복제 시간이 많이 소요됨
- 결과 도출을 위한 시간 소모가 큼
- Real-Time 데이터에 대한 분석의 어려움

Scale-Out Data Lake Analytics Strategy with Isilon

- 고립된 데이터 및 분석을 단일의 Shared Storage Infrastructure에 통합
  - TCO 절감과 함께 효율성 향상 및 결과 도출을 위한 시간 단축
- 급속한 변화속에서 투자 보호 제공
  - Optimally scale, End data migrations, Reuse assets
- "Enterprise data hub" / "Data lake"
  - 검증된 Enterprise 기능들을 Big Data 분석에 적용
- 업계 선도의 분석 ISV들과의 연합 솔루션 추진
  - Pivotal, Cloudera, Hortonworks, Splunk, RainStor, SAS Grid

EMC²

Scale-Out Data Lake Foundation




EMC²