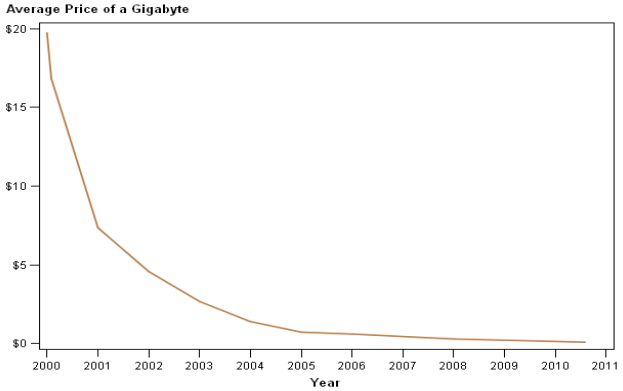
Mis on BIG DATA ja kuidas seda töödelda

Oleg Bogdanov
SAS Institute,
tehniline konsultant

Analüütika väljakutsed sel aastakümnel

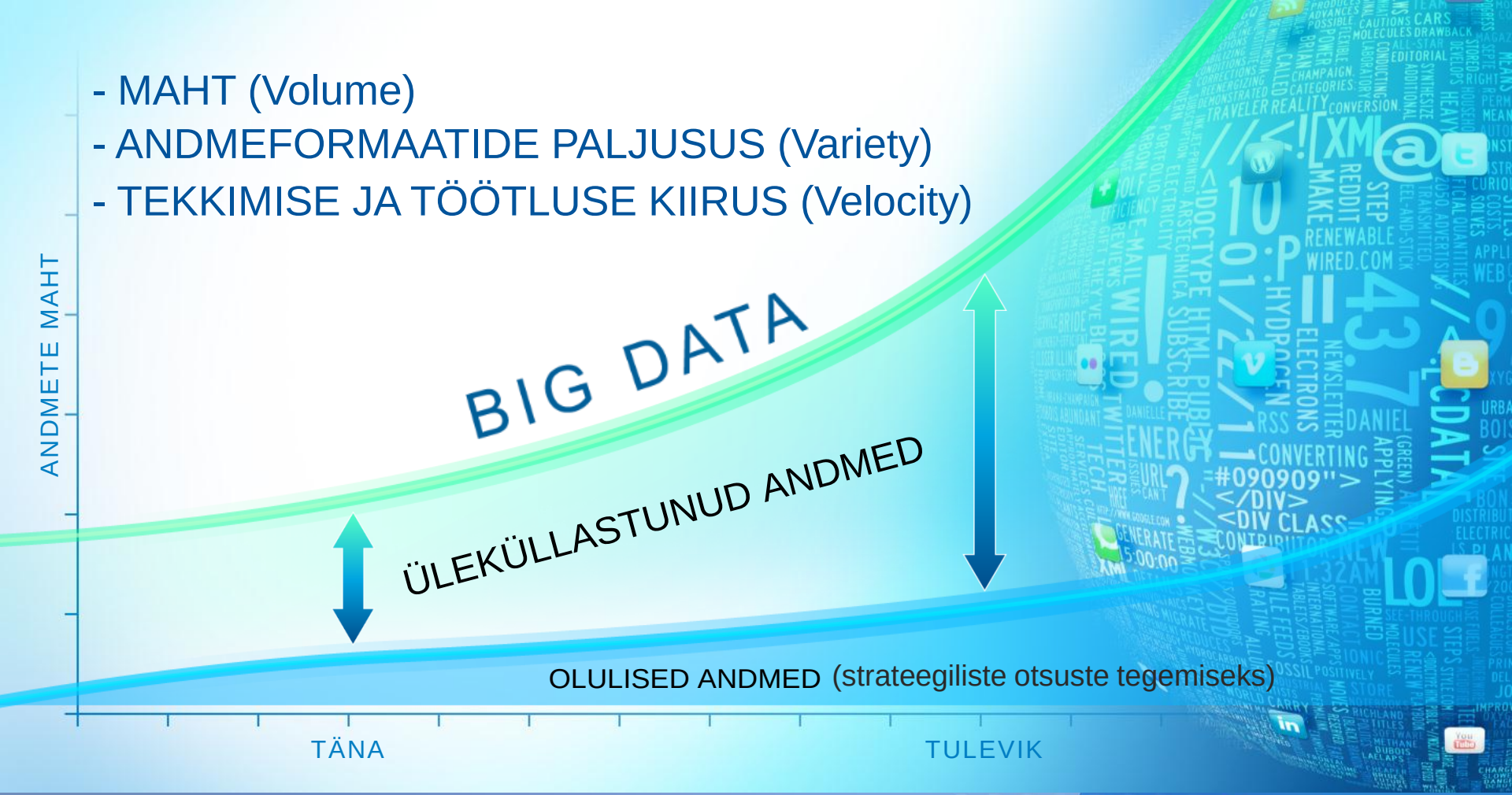
- Andmesalvestuse hinna jätkuv langus
  - Aastal 2000 1GB maksis keskmiselt $16.06,

    1 TB Andmeait oli suht haruldane nähtus
  - Nüüd maksab GB kettamälu $0.0621 ja TB alla $100



- Teiselt poolt tehnoloogia võimaldab
  - koguda ülisuuri andmemahtusid, infot objektide käitumisest, transaktsioonidest, harjumustest, tegevustest. Eilegi mainitud sotsiaalvõrgustike andmed, astronoomilised andmed

- MAHT (Volume)
- ANDMEFORMAATIDE PALJUSUS (Variety)
- TEKKIMISE JA TÖÖTLUSE KIIRUS (Velocity)



“Big Data” – uus lähenemine

See kõik eeldab ka analüütikult veidi teistmoodi lähenemist....

- Pigem heuristiline kui algoritmiline, andmetest lähtuv analüüs
- Andmete eelvaatlemise ja eelanalüüsi (data exploration) tähtsus

Vana hea normaliseeritud „Andmeait“ (EDW) peab muutuma „Analüütiliseks Andmeaidaks“ (ADW), mis ei ole enam täiesti normaliseeritud, kuid on orienteeritud analüütiliste ülesannete lahendamisele (mudelid)

„Big Data“ on suured mahud struktureeritud ja struktureerimata andmeid, mille haldamine tavapäraste relatsioonilise andmebaasi- ja andmetöötluse vahenditega on raskendatud, kui mitte võimatu (maht, formaadid, kiirus) .

Mis teha ?

Uute riistvaraliste vahendite kasutuselevõtt (kiiremad protsessorid, SSD)

Andmetöötlus jagamine paralleelselt käivitatavateks mooduliteks ja vastavate algoritmide arendamine

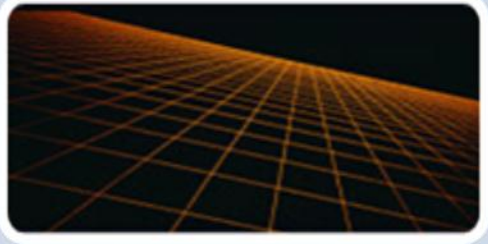
Andmed ja töötlemine peavad olema võimalikult lähestikku.

- Siirdada töötlemine andmete juurde
- Tuua andmed mälusse, kus toimub töötlemine

Võimalikud kombinatsioonid ülalmainitutest (palju sõltub sellest, mis kujul andmed on)

Kolm tehnoloogiat BIG DATA töötlemiseks

SAS® High-Performance Computing



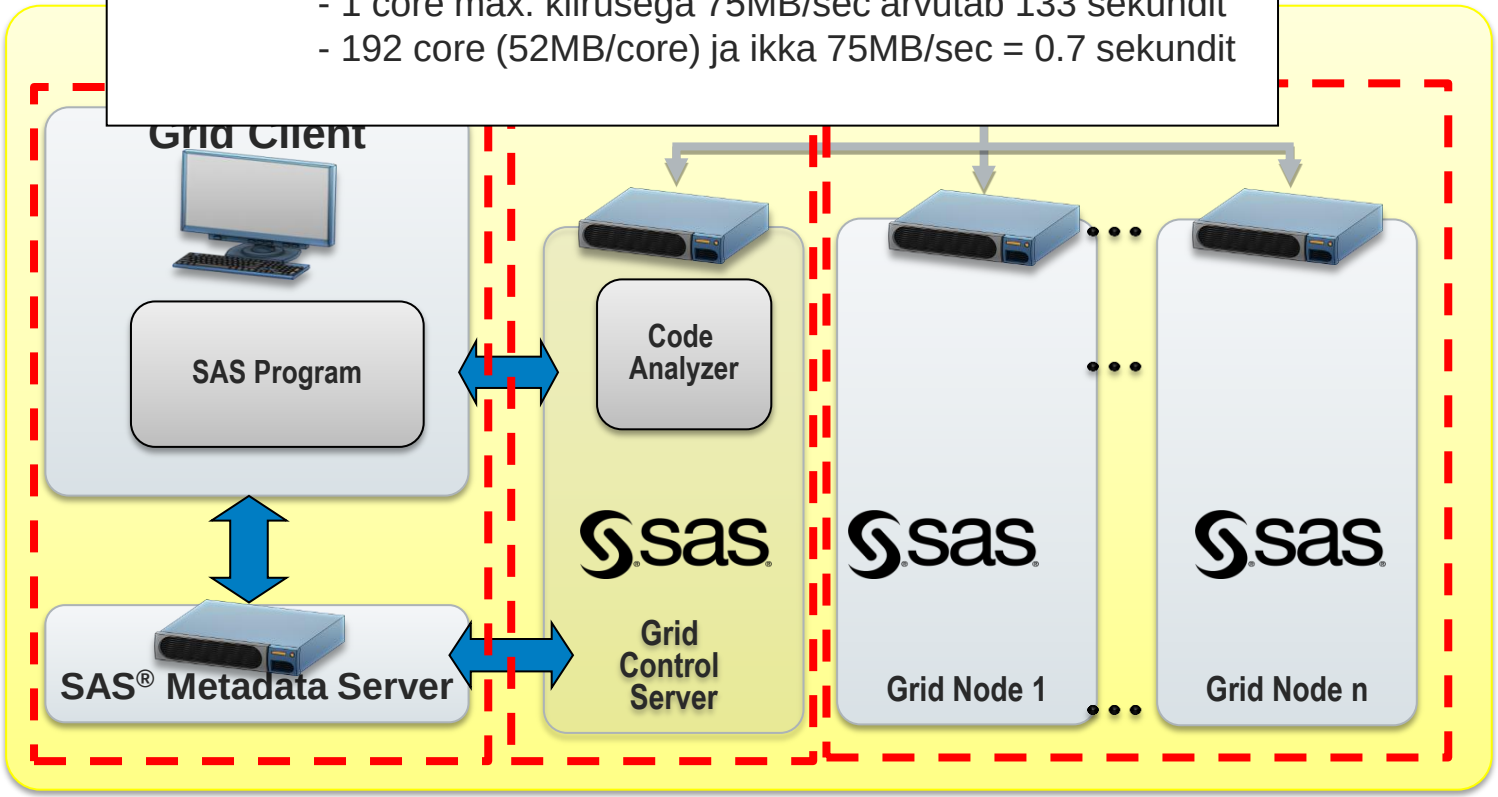
SAS® Grid Computing

SAS® In-Database

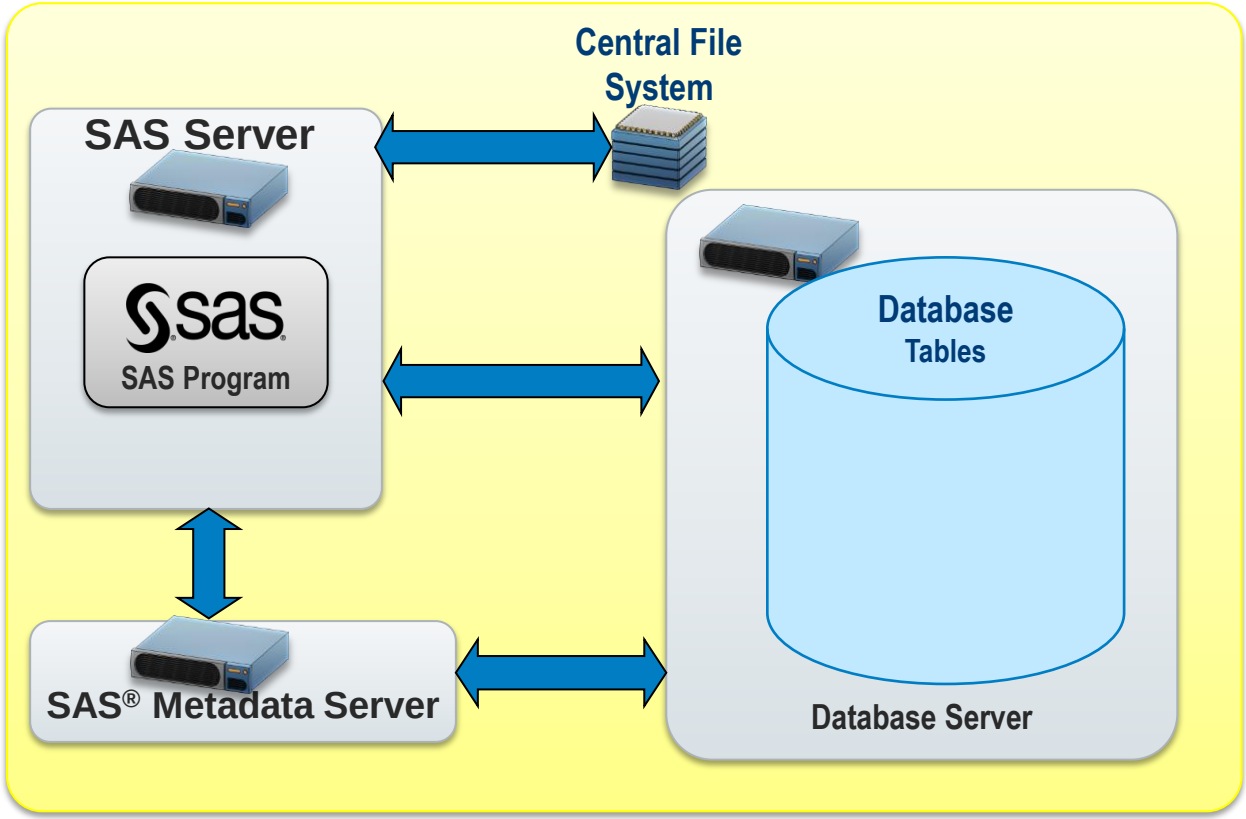
SAS® In-Memory Analytics

SAS Grid tehnoloogia

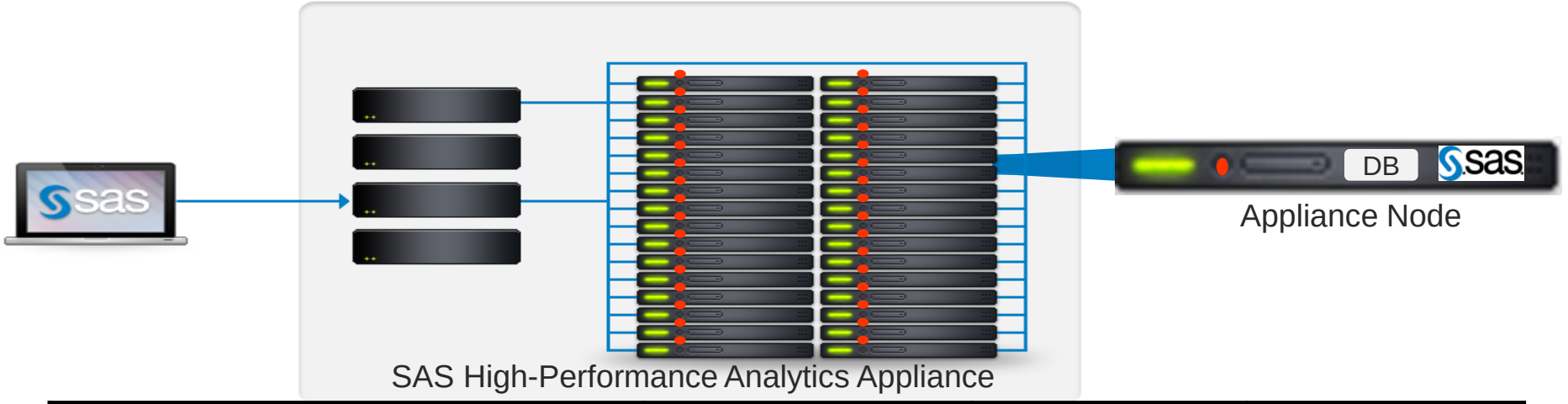
10GB andmetabel, mida oleks vaja töödelda
- 1 core max. kiirusega 75MB/sec arvutab 133 sekundit
- 192 core (52MB/core) ja ikka 75MB/sec = 0.7 sekundit



SAS In-Database



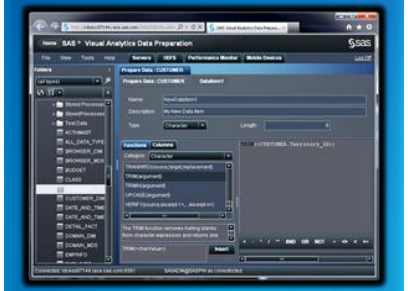
SAS High-Performance Analytics (Appliance)



Probleem	*Andmed*	*Enne*	*SAS HPA*
Laenutagastuse tõenäosuse skoorimine	1 miljard rida	11-20 tundi	54 sekundit
Müügikampaania vastuste genereerimine kontaktiajaloo põhjal	100M rida kontaktiajalugu 15M klienti	2,5 kuni 5 tundi	90 sekundit

SAS LASR Analytics Server technology

Central Entry Point | Integration | Role-based Views



DATA PREPARATION

- Monitor SAS® LASR™ Analytic server
- Load and join data
- Create calculated columns

EXPLORER

- Perform ad-hoc analysis and data discovery

DESIGNER

- Create dashboard style reports for web or mobile

MOBILE BI

- Native iOS application that delivers interactive reports created in the designer

SAS® LASR™ ANALYTIC SERVER

Apache HADOOP

Aitäh

Oleg.Bogdanov@sas.com

www.sas.com