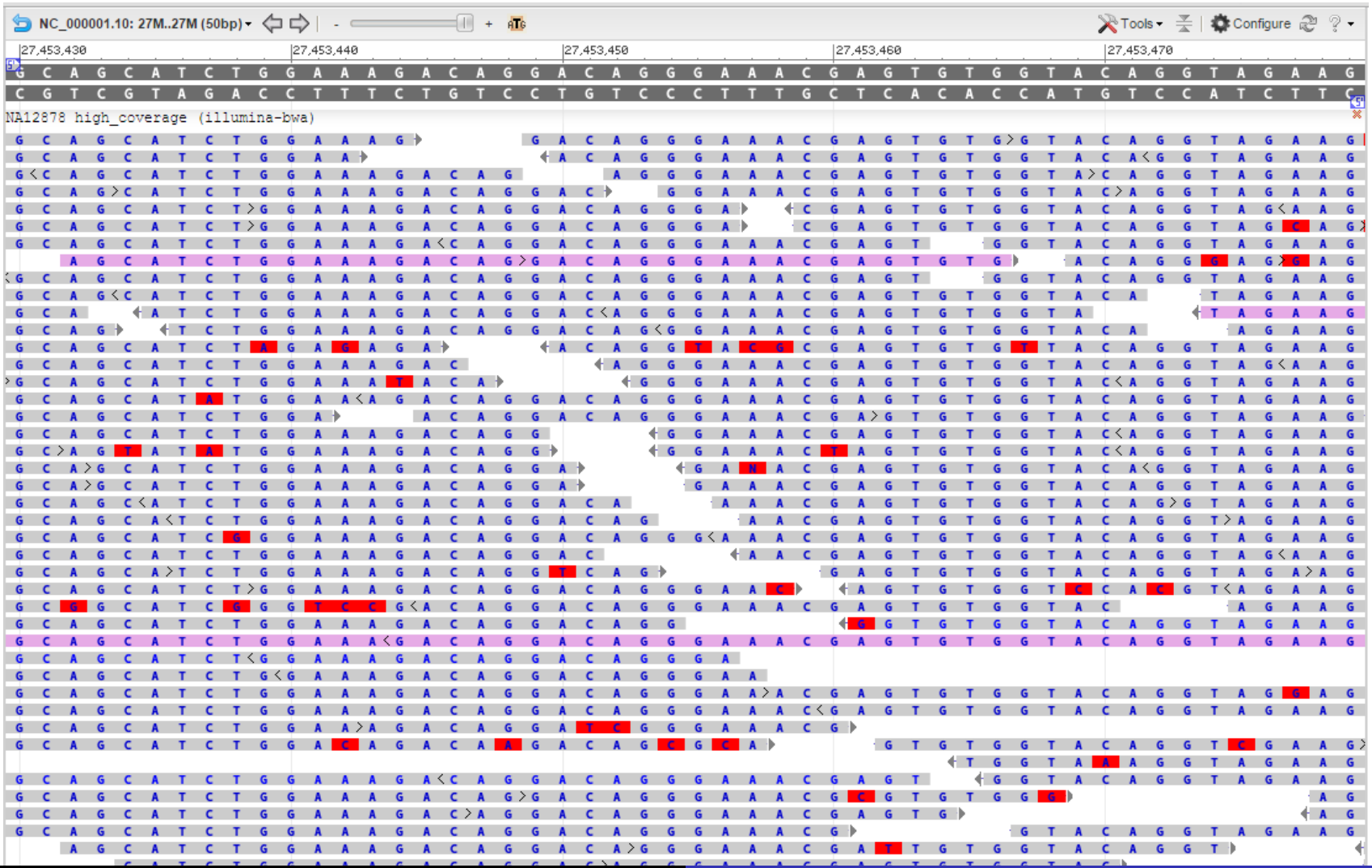
Custom Tooling for Loading Petabytes of Genomic Data into SciDB

Douglas J. Slotta

Stanford, CA
May 20th, 2015

NCBI

First We Chop You Up, Then We Put You Back Together

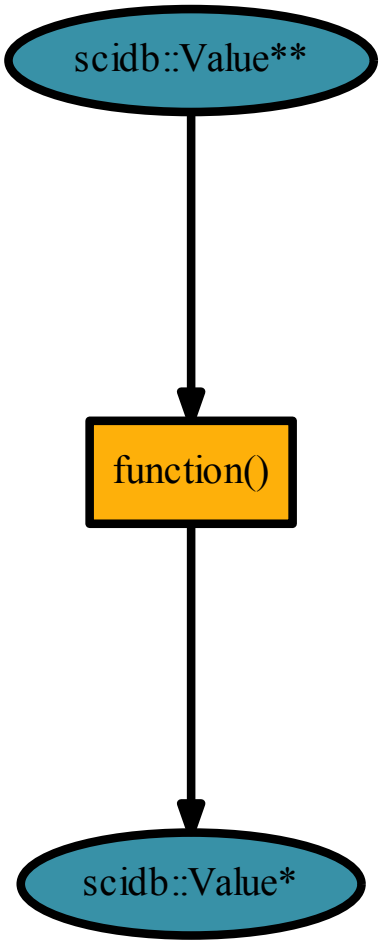


Attributes, Dimensions, and Datatypes, Oh My!

Attributes											
Name	ref	cov	A	C	G	T	D	Ia	Ic	Ig	Ig
Datatype	char	uint32	uint32	uint32	uint32	uint32	uint32	uint32	uint32	uint32	uint32

Dimensions			
Name	chrom	pos	sample
Range/Chunksize	1:25 / 1	1:250,000,000 / 25,000	1:* / 500

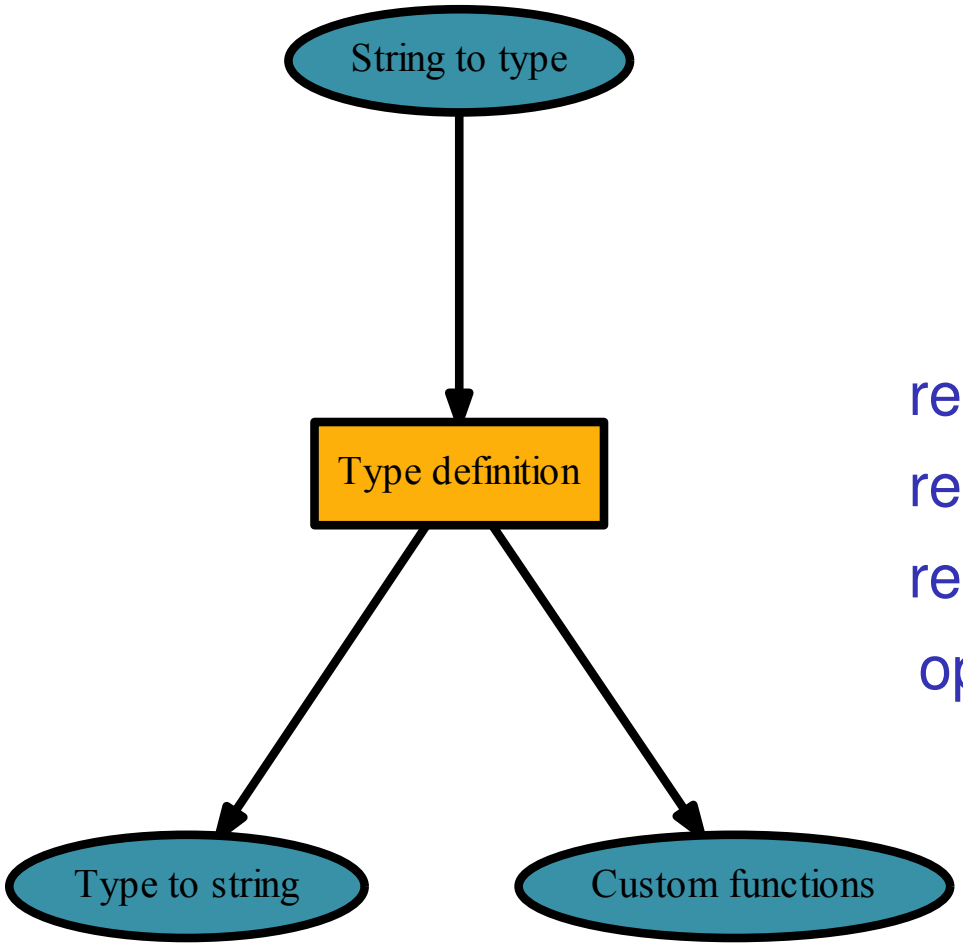
The Function Two-Step



input Array of scidb::Values

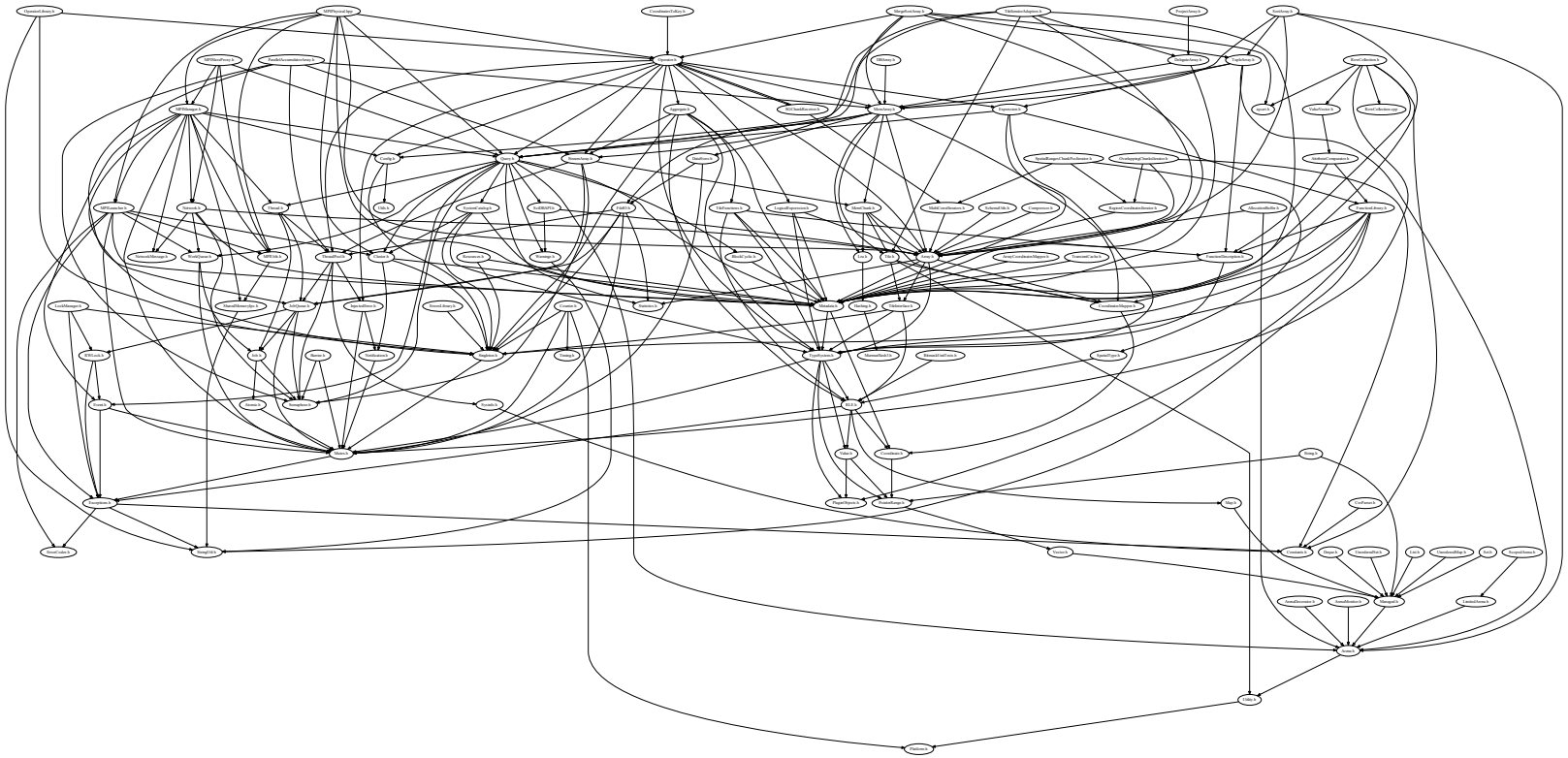
output Single scidb::Value

Forging New Datatypes



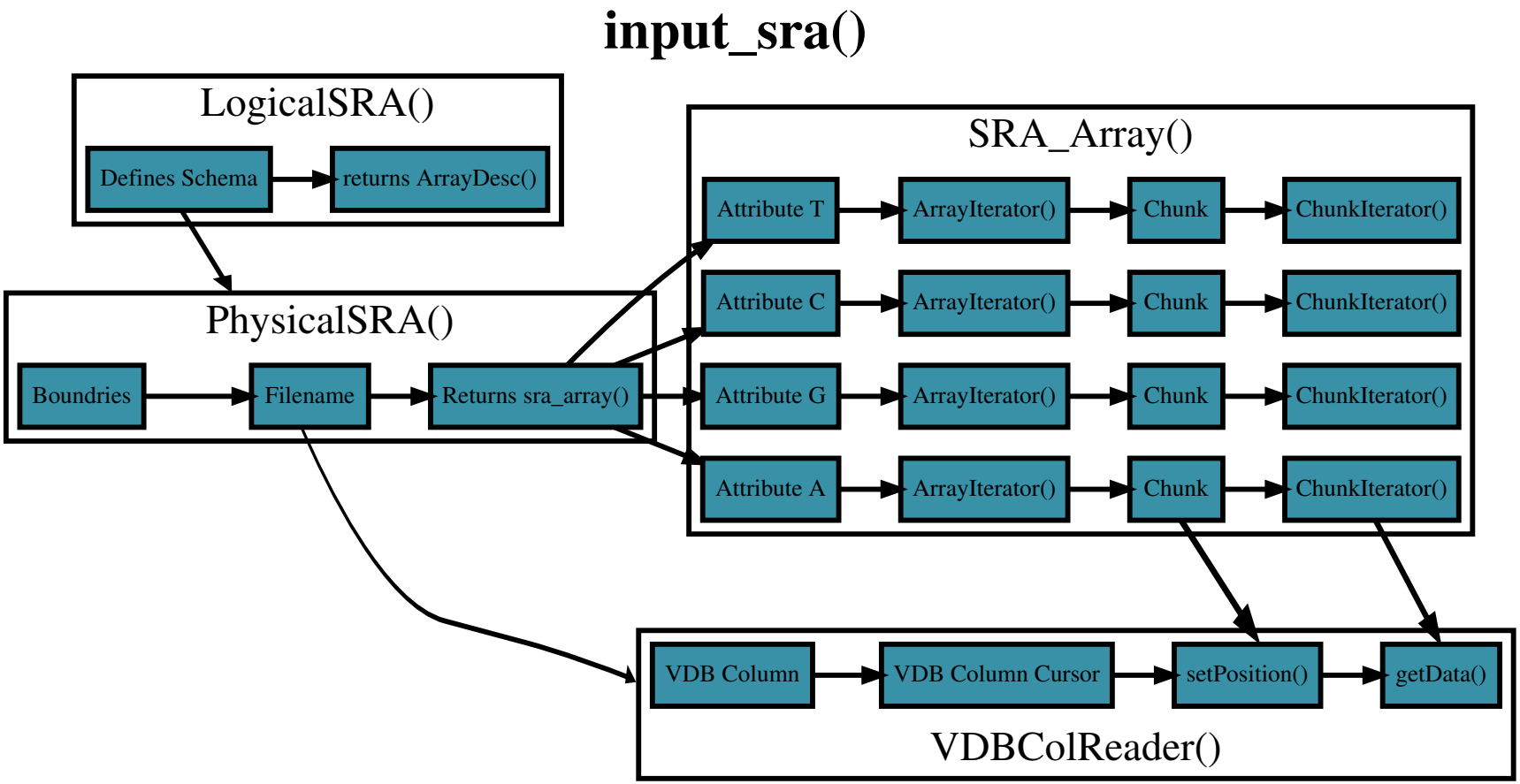
required	Size in bits
required	toString function
required	fromString function
optional	other functions

To Make an Apple Pie From Scratch You Must First Invent the Universe



NCBI

Embrace, Extend (& Extinguish?)



For a Cool Quarter Million . . .

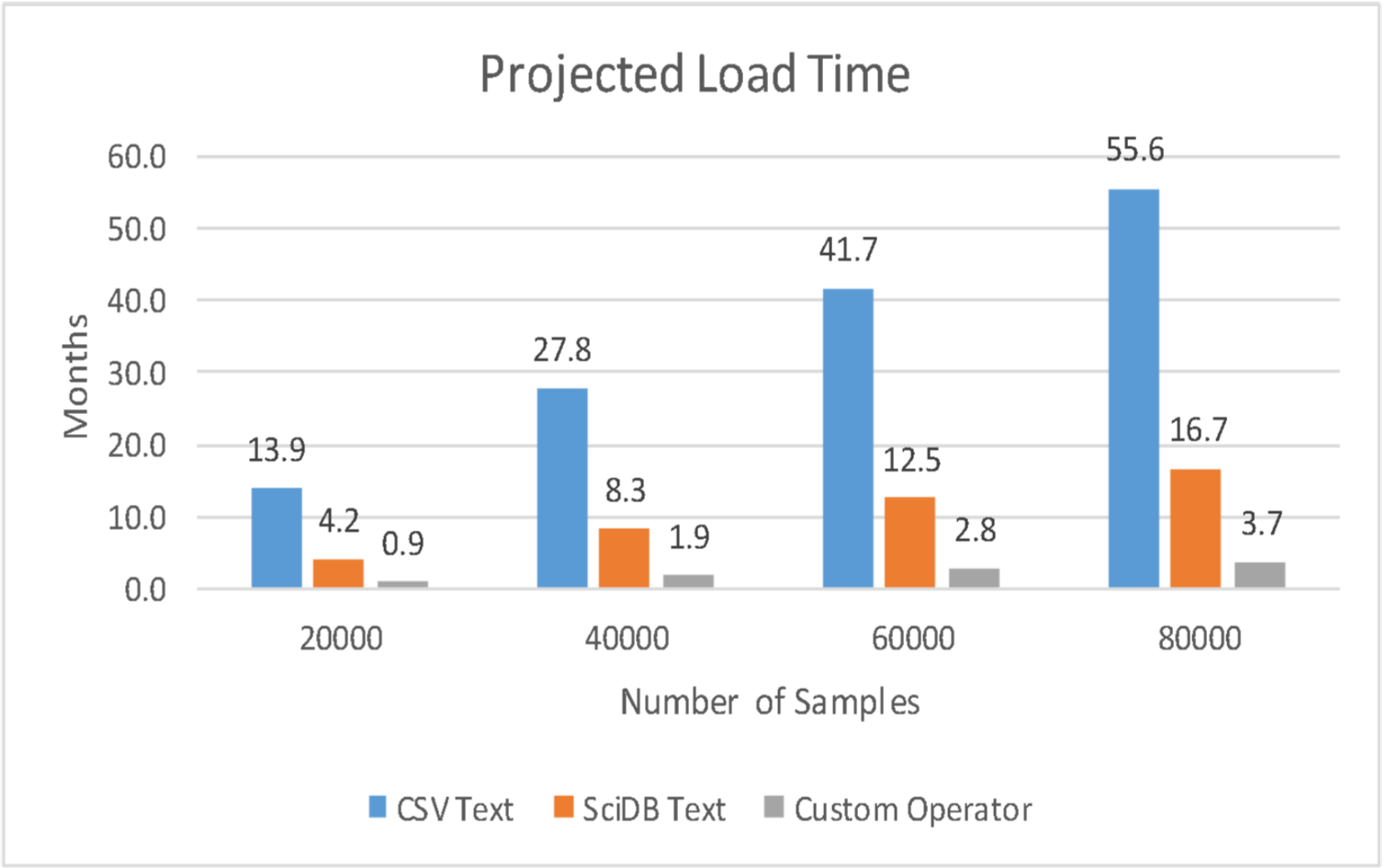
SciDB Cluster: 12 Dell R720s		
	Per node	Cluster total
RAM	128GB	1.5TB
Hard Disk	37TB (raid 6)	444TB
Cores	16 Xeon E5-2670	192 Xeon E5-2670
SciDB Instances	8	96

Slow, Less Slow, Least Slowest!

Load Throughput	
CSV Text, w/redimension()	30 min/sample
SciDB Text, parallel load	9 min/sample
Custom VDB load operator	2 min/sample

NCBI

Saving Years of Work



Other Considerations

Comparative Data Size/Query Speed		
	Disk Footprint	Query Speed
VDB	227 MB/sample	120 samples/second
SciDB	11.8 GB/sample	1,360 samples/second

Acknowledgments

Douglas Slotta	Lead Developer
Charlie Liu	Developer
Marty Brandon	Developer
Morty Abzug	Systems Admin
Don Preuss	Systems Facilitator
Alex Poliakov	SciDB Guru

Advice and Consent:

Steve Sherry

Jim Ostell

Karl Sirotkin

Eugene Yaschenko

Kurt Rodarmer