Symbolic Query Exploration

Pavel Grigorenko (joint work with Margus Veanes, Peli de Halleux and Nikolai Tillmann)

Institute of Cybernetics
Tallinn University of Technology

Theory Days at Mäetaguse 2 October 2009

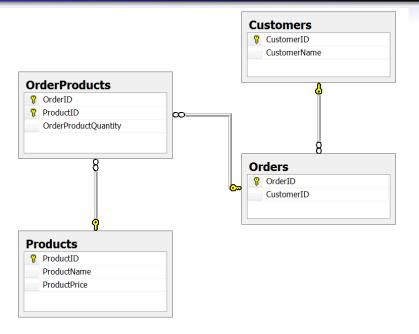


Overview

- Motivation
 - Test data generation for SQL queries
 - Parameter generation for queries and store procedures
- Goal
 - Investigate model generation with Z3
- Approach
 - ullet Satisfiability modulo background theory \mathcal{T}^Σ
 - ullet Mapping SQL to \mathcal{T}^Σ



Sample DB





Query

```
SELECT C.CustomerID, SUM(OP.OrderProductQuantity

* P.ProductPrice)

FROM OrderProducts AS OP

JOIN Orders AS O ON OP.OrderID = O.OrderID

JOIN Products AS P ON OP.ProductID = P.ProductID

JOIN Customers AS C ON O.CustomerID = C.CustomerID

WHERE @value > 1

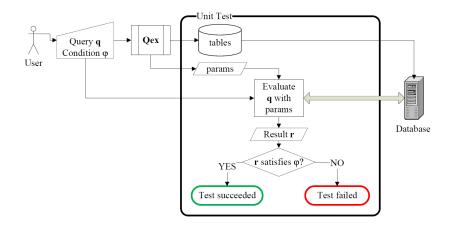
GROUP BY C.CustomerID

HAVING SUM(OP.OrderProductQuantity

* P.ProductPrice) > 100 + @value
```

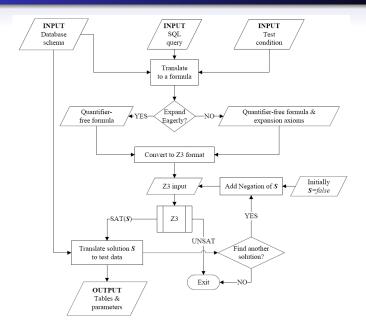


Unit test





Qex internal workflow





Expressions in \mathcal{T}^Σ

$$T^{\sigma} \qquad ::= x^{\sigma} \mid Default^{\sigma} \mid Ite(T^{\mathbb{B}}, T^{\sigma}, T^{\sigma}) \mid TheElementOf(T^{\mathbb{S}(\sigma)}) \mid \pi_{i}(T^{\mathbb{T}(\sigma_{0}, ..., \sigma_{i} = \sigma, ...)})$$

$$T^{\mathbb{T}(\sigma_{0}, ..., \sigma_{k})} ::= \langle T^{\sigma_{0}}, ..., T^{\sigma_{k}} \rangle$$

$$T^{\mathbb{Z}} ::= k \mid T^{\mathbb{Z}} + T^{\mathbb{Z}} \mid k * T^{\mathbb{Z}} \mid \Sigma_{i}(T^{\mathbb{S}(\mathbb{T}(\sigma_{0}, ..., \sigma_{i} = \mathbb{Z}, ...))})$$

$$T^{\mathbb{R}} ::= r \mid T^{\mathbb{R}} + T^{\mathbb{R}} \mid k * T^{\mathbb{R}} \mid \Sigma_{i}(T^{\mathbb{S}(\mathbb{T}(\sigma_{0}, ..., \sigma_{i} = \mathbb{R}, ...))}) \mid AsReal(T^{\mathbb{Z}})$$

$$T^{\mathbb{B}} ::= true \mid false \mid \neg T^{\mathbb{B}} \mid T^{\mathbb{B}} \wedge T^{\mathbb{B}} \mid T^{\mathbb{B}} \vee T^{\mathbb{B}} \mid T^{\sigma} = T^{\sigma} \mid T^{\mathbb{S}(\sigma)} \subseteq T^{\mathbb{S}(\sigma)} \mid T^{\sigma} \in T^{\mathbb{S}(\sigma)} \mid T^{\mathbb{Z}} \leq T^{\mathbb{Z}} \mid T^{\mathbb{R}} \leq T^{\mathbb{R}}$$

$$T^{\mathbb{S}(\sigma)} ::= X^{\mathbb{S}(\sigma)} \mid \{T^{\sigma} \mid_{\bar{x}} T^{\mathbb{B}}\} \mid T^{\mathbb{S}(\sigma)} \cup T^{\mathbb{S}(\sigma)} \mid T^{\mathbb{S}(\sigma)} \cap T^{\mathbb{S}(\sigma)} \mid T^{\mathbb{S}(\sigma)} \setminus T^{\mathbb{S}(\sigma)} \setminus T^{\mathbb{S}(\sigma)} \setminus T^{\mathbb{S}(\sigma)}$$

$$F ::= T^{\mathbb{B}} \mid \exists x F \mid \exists x F \mid \exists x F \mid T^{\mathbb{S}(\sigma)} \in T^{\mathbb{S}(\sigma)} \mid T^{\mathbb{S}(\sigma)} \cap T^{\mathbb{S}($$



- S is a state for a term t such that $FV(t) \subseteq Dom(S)$
- For a given term t, t^S is the *interpretation* of t in S
- For a given formula φ , $S \models \varphi$ means that φ^S is *true*

$$Ite(\varphi, t_1, t_2)^S = \begin{cases} t_1^S, & \text{if } S \models \varphi; \\ t_2^S, & \text{otherwise.} \end{cases}$$
$$\{t_0 \mid_{x^{\sigma}} \varphi\}^S = \{t_0^{S \uplus \{x \mapsto a\}} : a \in \mathcal{U}^{\sigma}, S \uplus \{x \mapsto a\} \models \varphi\}$$
$$\Sigma_i(t_1)^S = \sum_{a \in t_1^S} \pi_i(a)$$



- S is a state for a term t such that $FV(t) \subseteq Dom(S)$
- For a given term t, t^S is the *interpretation* of t in S
- For a given formula φ , $S \models \varphi$ means that φ^S is true

$$Ite(\varphi, t_1, t_2)^{S} = \begin{cases} t_1^{S}, & \text{if } S \models \varphi; \\ t_2^{S}, & \text{otherwise.} \end{cases}$$
$$\{t_0 \mid_{x^{\sigma}} \varphi\}^{S} = \{t_0^{S \uplus \{x \mapsto a\}} : a \in \mathcal{U}^{\sigma}, S \uplus \{x \mapsto a\} \models \varphi\}$$
$$\Sigma_i(t_1)^{S} = \sum_{a \in t_1^{S}} \pi_i(a)$$

Example

Integer multiplication:

$$n*m \stackrel{\mathrm{def}}{=} \Sigma_0(\{\langle m,x\rangle \mid 0 \leq x < n\}) = \sum_{x=0}^{n-1} \pi_0(\langle m,x\rangle) = \sum_{x=0}^{n-1} m$$



Tables are bags (multisets)!

A bag b with elements $\{a_i\}_{i< n}$ each having multiplicity $m_i>0$ in b for i< n, is represented as a set of pairs $\{\langle a_i,m_i\rangle\}_{i< n}$, thus having the sort $\mathbb{S}(\mathbb{T}(\sigma,\mathbb{Z}))$ for some basic sort σ (domain sort of b). Let $\mathbb{M}(\sigma)$ be the type $\mathbb{S}(\mathbb{T}(\sigma,\mathbb{Z}^+))$ with the constraint:

$$\forall X^{\mathbb{M}(\sigma)} \,\forall x^{\sigma} \, y^{\sigma} \, \big(\big(x \in X \wedge y \in X \wedge x.0 = y.0 \big) \Rightarrow x.1 = y.1 \big).$$



Tables are bags (multisets)!

A bag b with elements $\{a_i\}_{i< n}$ each having multiplicity $m_i>0$ in b for i< n, is represented as a set of pairs $\{\langle a_i,m_i\rangle\}_{i< n}$, thus having the sort $\mathbb{S}(\mathbb{T}(\sigma,\mathbb{Z}))$ for some basic sort σ (domain sort of b). Let $\mathbb{M}(\sigma)$ be the type $\mathbb{S}(\mathbb{T}(\sigma,\mathbb{Z}^+))$ with the constraint:

$$\forall X^{\mathbb{M}(\sigma)} \,\forall x^{\sigma} \, y^{\sigma} \, \big(\big(x \in X \land y \in X \land x.0 = y.0 \big) \Rightarrow x.1 = y.1 \big).$$

$$\begin{array}{ccc} \textit{AsBag}(\textit{Y}^{\mathbb{S}(\sigma)}) & \stackrel{\text{def}}{=} & \{\langle y, 1 \rangle \mid y \in \textit{Y}\} \\ \textit{AsSet}(\textit{X}^{\mathbb{M}(\sigma)}) & \stackrel{\text{def}}{=} & \{y.0 \mid y \in \textit{X}\} \\ \Sigma^{\text{b}}_{i}(\textit{X}^{\mathbb{M}(\mathbb{T}(\sigma_{0}, ..., \sigma_{i}, ...))}) & \stackrel{\text{def}}{=} & \Sigma_{0}(\{\langle x.1 * x.0.i, x.0 \rangle \mid x \in \textit{X}\}) \end{array}$$



Example

$$\begin{split} \varphi[x^{\mathbb{Z}}] &= x < 4 \\ q[X^{\mathbb{M}(\mathbb{T}(\mathbb{Z},\mathbb{Z},\mathbb{Z}))}] &= \{\langle x.0.0, \Sigma_1^{\mathrm{b}}(\{y \mid y \in X \land x.0.0 = y.0.0 \land \varphi[y.0.2]\}) \rangle \\ &\quad | x \in X \land \varphi[x.0.2] \} \\ t &= \{\langle \langle 0, 2, 1 \rangle, 2 \rangle, \langle \langle 1, 2, 3 \rangle, 1 \rangle, \langle \langle 1, 2, 4 \rangle, 1 \rangle \} \end{split}$$

$$q[t] &= \{\langle x.0.0, \Sigma_1^{\mathrm{b}}(\{y \mid y \in t \land x.0.0 = y.0.0 \land \varphi[y.0.2]\}) \rangle \\ &\quad | x \in t \land \varphi[x.0.2] \} \\ &= \{\langle 0, \Sigma_1^{\mathrm{b}}(\{y \mid y \in t \land 0 = y.0.0 \land \varphi[y.0.2]\}) \rangle, \\ &\quad \langle 1, \Sigma_1^{\mathrm{b}}(\{y \mid y \in t \land 1 = y.0.0 \land \varphi[y.0.2]\}) \rangle \} \\ &= \{\langle 0, \sum_{a \in \{\langle \langle 0, 2, 1 \rangle, 2 \rangle\}} \pi_1(a) * \pi_1(\pi_0(a)) \rangle, \\ &\quad \langle 1, \sum_{a \in \{\langle \langle 1, 2, 3 \rangle, 1 \rangle\}} \pi_1(a) * \pi_1(\pi_0(a)) \rangle \} \\ &= \{\langle 0, 4 \rangle, \langle 1, 2 \rangle \} \end{split}$$



Translation

 $\mathbf{Q}: \mathrm{SQL} \to \mathcal{T}^{\Sigma}$

Strings

- All strings have a maximum length k; an encoding of a k-string is as a k-tuple of integers, each character a is encoded as an integer c(a). For a string $a_0 \cdots a_l$, l < k the encoding is $\langle c(a_0), \ldots, c(a_l), 0, \ldots, 0 \rangle$.
- For a collection D of strings, encode as |D|-enums.

Nullable values

Given a basic sort σ , let $?\sigma$ be the sort $\mathbb{T}(\sigma,\mathbb{B})$ with the constraint

 $\forall x^{?\sigma} (x.1 = \textit{false} \Rightarrow x.0 = \textit{Default}^{\sigma}) \text{ and } \textit{null}^{?\sigma} \stackrel{\text{def}}{=} \textit{Default}^{\mathbb{T}(\sigma, \mathbb{B})}.$

Operations that are defined for σ are lifted to $?\sigma$. For example, for a numeric sort σ ,

$$x^{?\sigma} + y^{?\sigma} \stackrel{\text{def}}{=} lte(x.1 \land y.1, \langle x.0 + y.0, true \rangle, null^{?\sigma}).$$



Select clauses

$$\begin{aligned} \mathbf{Q}(\text{SELECT 1 FROM t}) & \stackrel{\text{def}}{=} \left\{ \left\langle \left\langle x.0.l_0, \ldots, x.0.l_n \right\rangle, M(x) \right\rangle \mid x \in \mathbf{Q}(\texttt{t}) \right\} \\ & \text{where } M(x) = \Sigma_0(\left\{ \left\langle y.1, y \right\rangle \mid y \in \mathbf{Q}(\texttt{t}) \land \bigwedge_{i=0}^n y.0.l_i = x.0.l_i \right\}) \end{aligned}$$

$$\mathbf{Q}(\text{SELECT DISTINCT 1 FROM t}) & \stackrel{\text{def}}{=} \\ & AsBag(AsSet(\mathbf{Q}(\text{SELECT 1 FROM t})))$$

$$AsSet(\mathbf{Q}(\text{SELECT 1 FROM t})) = \end{aligned}$$

 $\{\langle y.l_0,\ldots,y.l_n\rangle\mid y\in AsSet(\mathbf{Q}(t))\}$



Join clauses

$$\begin{aligned} & \mathbf{Q}(\texttt{t1 INNER JOIN t2 ON c}) \stackrel{\text{def}}{=} \\ & \{ \langle x_1.0 \times x_2.0, x_1.1 * x_2.1 \rangle \mid x_1 \in \mathbf{Q}(\texttt{t1}) \land x_2 \in \mathbf{Q}(\texttt{t2}) \land \mathbf{Q}(\texttt{c})[x_1.0, x_2.0] \} \\ & x \times y \stackrel{\text{def}}{=} \langle \pi_0(x), \dots, \pi_{m-1}(x), \pi_0(y), \dots, \pi_{n-1}(y) \rangle \\ & AsSet(\mathbf{Q}(\texttt{t1 INNER JOIN t2 ON c})) = \\ & \{ y_1 \times y_2 \mid y_1 \in AsSet(\mathbf{Q}(\texttt{t1})) \land y_2 \in AsSet(\mathbf{Q}(\texttt{t2})) \land \mathbf{Q}(\texttt{c})[y_1, y_2] \} \end{aligned}$$



Grouping and aggregates

```
\begin{split} &\texttt{t} = \texttt{SELECT a, SUM(b)} \  \  \, \texttt{AS d FROM t1 WHERE c1} \\ &\textbf{Q}(\texttt{t GROUP BY a HAVING c2}) \stackrel{\text{def}}{=} \textit{AsBag}(\{z \mid z \in \textit{G} \land \textbf{Q}(\texttt{c2})[z]\}) \\ &\text{where } \textit{G} = \{\langle x.0.0, \Sigma_1^b(\{y \mid y \in \textbf{Q}(\texttt{t}) \land y.0.0 = x.0.0\})\rangle \mid x \in \textbf{Q}(\texttt{t})\} \\ &\textit{Sum} \stackrel{\text{def}}{=} \Sigma_i^b \\ &\textit{Count} \stackrel{\text{def}}{=} \Sigma_1 \\ &\textit{Min}(\textit{X}^{\mathbb{S}(\sigma)}) \stackrel{\text{def}}{=} \textit{TheElementOf}(\{y \mid y \in \textit{X} \land \{z \mid z \in \textit{X} \land z < y\} = \emptyset\}) \end{split}
```



Union

$$\mathbf{Q}(q1 \text{ UNION } q2) \stackrel{\text{def}}{=} AsBag(AsSet(\mathbf{Q}(q1)) \cup AsSet(\mathbf{Q}(q2))).$$



Simplifications

$$\begin{array}{lll} \textit{AsSet}(\textit{AsBag}(\textit{X}^{\mathbb{S}(\sigma)})) & = & \textit{X} \\ & \Sigma_{i}^{\text{b}}(\textit{AsBag}(\textit{X}^{\mathbb{S}(\sigma)})) & = & \Sigma_{i}(\textit{X}) \\ & \textit{AsSet}(\{t \mid \varphi\}^{\mathbb{M}(\sigma)}) & = & \{t.0 \mid \varphi\} \\ & \pi_{i}(\langle t_{0}, \ldots, t_{i}, \ldots \rangle) & = & t_{i} \end{array}$$



Model Generation with SMT

Model Generation in $\mathcal{T}_{\mathbf{Q}}^{\Sigma}$

Given a quantifier free formula $\varphi[X]$ in \mathcal{T}^{Σ} , and a query q, decide if $\psi = \varphi[\mathbf{Q}(q)]$ is satisfiable, and if ψ is satisfiable, generate a model of ψ .

Two approaches to deal with comprehensions and summations in Z3

- Eager expansion
 - Provide finite bounds for tables and unwind
- Lazy expansion
 - · Add expansion rules as axioms



Expansion of terms

- Given a query q[X]
- ② Create a symbolic table $t_X = \{\langle x_1, m_1 \rangle, \dots, \langle x_k, m_k \rangle\}$ (k and m_i are fixed, x_i are variables)
- **3** Expand $q[t_X]$ to $\varphi = \mathbf{Exp}(q[t_X])$
- Generate a model for φ (if φ is sat.)
- **1..4.** Increase k and m_i , repeat 1..4.



Eager expansion

Set describer

Consider a formula $\psi[\overline{X}]$ as an instance of the model generation problem, where every X in \overline{X} is a bag variable.

- The constant $Empty^{\mathbb{S}(\sigma)}$ is a set describer.
- If $t^{\mathbb{S}(\sigma)}$ is a set describer then so is the term $Set(\varphi^{\mathbb{B}}, u^{\sigma}, t)$.

Given a state S for $Set(\varphi, u, t)$, the interpretation in S is,

$$Set(\varphi, u, t)^{S} = Ite(\varphi, \{u\}, \emptyset)^{S} \cup t^{S}, \quad Empty^{S} = \emptyset.$$

X is fixed in \overline{X} and t_X is the set describer:

$$Set(true, \langle x_1, m_1 \rangle, \dots Set(true, \langle x_k, m_k \rangle, Empty) \dots)$$

where k and all the m_i 's are some positive integer constants and each x_i is a variable.



Expansion rules

Expansion rule for *comprehensions*

$$\begin{aligned} & \operatorname{Exp}(\{t \mid_{X} x \in r \land \varphi\}) & \stackrel{\text{def}}{=} & \operatorname{ExpC}(t, x, \operatorname{Exp}(r), \varphi) \\ & \operatorname{ExpC}(t, x, \operatorname{Empty}, \varphi) & \stackrel{\text{def}}{=} & \operatorname{Empty} \end{aligned}$$

$$& \operatorname{ExpC}(t[x], x, \operatorname{Set}(\gamma, u, \operatorname{rest}), \varphi[x]) & \stackrel{\text{def}}{=} & \operatorname{Set}(\gamma \land \operatorname{Exp}(\varphi[u]), \operatorname{Exp}(t[u]), \\ & \operatorname{ExpC}(t, x, \operatorname{rest}, \varphi)) \end{aligned}$$



Expansion rules

Expansion rule for comprehensions

$$\begin{aligned} & \operatorname{Exp}(\{t \mid_{x} x \in r \wedge \varphi\}) & \stackrel{\text{def}}{=} & \operatorname{ExpC}(t, x, \operatorname{Exp}(r), \varphi) \\ & \operatorname{ExpC}(t, x, \operatorname{Empty}, \varphi) & \stackrel{\text{def}}{=} & \operatorname{Empty} \end{aligned}$$

$$& \operatorname{ExpC}(t[x], x, \operatorname{Set}(\gamma, u, \operatorname{rest}), \varphi[x]) & \stackrel{\text{def}}{=} & \operatorname{Set}(\gamma \wedge \operatorname{Exp}(\varphi[u]), \operatorname{Exp}(t[u]), \\ & \operatorname{ExpC}(t, x, \operatorname{rest}, \varphi)) \end{aligned}$$

Expansion rule for Σ_i

$$egin{array}{lll} \mathbf{Exp}(\Sigma_i(t)) &\stackrel{ ext{def}}{=} & \mathbf{Sum}_i(\mathbf{Exp}(t), \mathit{Empty}) \ & \mathbf{Sum}_i(\mathit{Empty}, s) &\stackrel{ ext{def}}{=} & 0 \ & \mathbf{Sum}_i(\mathit{Set}(\gamma, u, \mathit{rest}), s) &\stackrel{ ext{def}}{=} & \mathit{Ite}(\gamma \wedge u \notin s, \pi_i(u), 0) & + \ & + \mathbf{Sum}_i(\mathit{rest}, \mathit{Set}(\gamma, u, s)) \end{array}$$



Lazy expansion

In addition to a quantifier free formula ψ , universally quantified *axioms* are provided in the form:

$$(\forall \bar{x}(\alpha), pat_{\alpha}), FV(\alpha) = FV(pat_{\alpha}) = \bar{x}$$

If ψ contains a subterm t and there exists a substitution θ such that $t=pat_{\alpha}\theta$, i.e., t matches the pattern pat_{α} , then ψ is replaced during proof search by (a reduction of) $\psi \wedge \alpha \theta$.



Lazy sum

Axioms

$$\begin{array}{rcl} \alpha_1 & = & \forall s (\mathbf{Sum}_i(\mathit{Empty}, s) = 0) \\ \mathit{pat}_{\alpha_1} & = & \mathbf{Sum}_i(\mathit{Empty}, s) \\ \alpha_2 & = & \forall b \, u \, r \, s (\mathbf{Sum}_i(\mathit{Set}(b, u, r), s) = \\ & & \mathit{Ite}(b \wedge u \notin s, \pi_i(u), 0) + \mathbf{Sum}_i(r, \mathit{Ite}(b, \{u\}, \emptyset) \cup s)) \\ \mathit{pat}_{\alpha_2} & = & \mathbf{Sum}_i(\mathit{Set}(b, u, r), s) \end{array}$$

Example

$$x \leq \mathbf{Sum}_{1}(Set(true, \langle 1, y \rangle, Set(true, \langle 1, z \rangle, Empty)), \emptyset)$$

$$\stackrel{\alpha_{2}}{\Longrightarrow} x \leq y + \mathbf{Sum}_{1}(Set(true, \langle 1, z \rangle, Empty), \{\langle 1, y \rangle\})$$

$$\stackrel{\alpha_{2}}{\Longrightarrow} x \leq y + Ite(z \neq y, z, 0) + \mathbf{Sum}_{1}(Empty, \{\langle 1, y \rangle, \langle 1, z \rangle\})$$

$$\stackrel{\alpha_{1}}{\Longrightarrow} x \leq y + Ite(z \neq y, z, 0)$$



Experiments

q1: SELECT C.CustomerID, O.OrderID FROM Orders AS O JOIN Customers AS C ON O.CustomerID = C.CustomerID WHERE O.CustomerID > 2 AND O.OrderID < 15 q2: SELECT C.CustomerID,
Count (0.0rderID)
FROM Orders AS O
JOIN Customers AS C ON
O.CustomerID = C.CustomerID
GROUP BY C.CustomerID HAVING
Count (0.0rderID) > 1

q3: DECLARE @value AS INT;

SELECT C.CustomerID, SUM(OP.OrderProductQuantity * P.ProductPrice)

FROM OrderProducts AS OP

JOIN Orders AS O ON OP.OrderID = O.OrderID

JOIN Products AS P ON OP.ProductID = P.ProductID

JOIN Customers AS C ON O.CustomerID = C.CustomerID

WHERE @value > 1

GROUP BY C.CustomerID

HAVING SUM(OP.OrderProductOuantity * P.ProductPrice) > 100 + @value

query	cond	k	check	$t_{ m exp}$	t_{z3}
q1	res ≠ ∅	1	sat	.03	.001
		2	sat	.05	.005
		3	sat	.3	.02
		4	sat	1.4	.13
	res = ∅	1	sat	.03	.001
		2	sat	.05	.006
		3	sat	.3	.12
		4	sat	1.4	2

query	cond	k	check	$t_{\rm exp}$	t_{z3}
q1		1	unsat	.03	.001
	res = 5	2	unsat	.05	.01
		3	unsat	.3	.16
		4	unsat	1.4	10
		5	sat	8.4	1.6
q2	$res \neq \emptyset$	1	unsat	.03	.001
		2	sat	.7	.006
		3	sat	26	.03
q3	$res eq \emptyset$	1	sat	.34	.001
		2	sat	30	.03
q3	$res \neq \emptyset$	1 2			

Implementation

- Qex written in C#, uses T-SQL syntax for input queries
- Integration with Pex



- Project homepage http://research.microsoft.com/qex
- Check Qex demo at http://channel9.msdn.com/posts/Peli/Qex-Symbolic-Query-Exploration/





Future work

- Lazy expansion (done?)
- Improved Database constraint support
- Queries with side-effects
- Improved datatype support
- Query optimizations



Future work

- Lazy expansion (done?)
- Improved Database constraint support
- Queries with side-effects
- Improved datatype support
- Query optimizations

Thanks! Questions?

